2017

Inference of Biogeographical Ancestry Under Resource Constraints

Tanjin Taher Toma

INFERENCE OF BIOGEOGRAPHICAL ANCESTRY UNDER RESOURCE CONSTRAINTS

Tanjin Taher Toma

Thesis submitted to the

Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in

Electrical Engineering

Donald Adjeroh, Ph.D., Chair

Jeremy Dawson, Ph.D.

YanFang Ye, Ph.D.

Tim Driscoll, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia

2017

Keywords: SNP; DNA; feature selection; ancestry prediction; single chromosome

ABSTRACT

INFERENCE OF BIOGEPGRAPHICAL ANCESTRY UNDER RESOURCE CONSTRAINTS

Tanjin Taher Toma

We study the problem of predicting human biogeographical ancestry using genomic data. While continental level ancestry prediction is relatively simple using genomic information, distinguishing between individuals from closely associated sub-populations (e.g., from the same continent) is still a difficult challenge. In particular, we focus on the case where the analysis is constrained to using single nucleotide polymorphisms (SNPs) from just one chromosome. We thus propose methods to construct ancestry informative SNP panels analyzing variants from a single chromosome, and evaluate the performance of such panels for both continental-level and sub-continental level ancestry prediction.

Efficient selection of ancestry informative SNPs is the key to successful ancestry prediction. The removal of redundant and noisy SNP features is essential prior to applying a learning algorithm. Here we propose two distinct methods of SNP selection: one is correlation-based SNP selection which uses a correlation metric to evaluate the usefulness of SNP features, while the other is random subspace projection based SNP selection which uses the learning algorithm itself to evaluate the worth of the SNP features. Correlation-based SNP selection approach can construct a small panel of useful SNPs for both continental level classification as well as binary classification of sub-populations. Unlike the correlation-based selection, random subspace projection based selection can construct efficient panel of SNP markers to address the difficult task of multinomial classification with multiple closely related sub-populations. We include results that demonstrate the performance of both methods, including comparison with other recently published related methods.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Donald Adjeroh, for his supervision throughout the course of my research. His continuous guidance and suggestions helped me to complete my thesis. Also, I greatly appreciate my thesis committee members, Dr. Jeremy Dawson, Dr. Yanfang Ye and Dr. Tim Driscoll, for their support and encouragement. I am also thankful to the members of my research group for their help during my research at WVU.

Besides, I wish to acknowledge the Lane Department of Computer Science and Electrical Engineering for giving me the opportunity to pursue my master's degree in such an academically vibrant environment.

Last but not the least, I would like to thank my family for supporting me throughout my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Background

Genomic ancestry inference is an active area of research in the field of bioinformatics, genetics, biomedical and forensic science. Accurate inference of genetic ancestry is useful for many purposes. For instance, population stratification can confound the relationship between a genetic marker and disease. Identifying ancestry informative markers (AIMs) in the genome is essentially useful for detecting such stratification in case-control association studies of complex diseases, such as diabetes, cardiovascular disease and cancer [1, 2, 3]. Measuring genetic ancestry has also been a focus in forensic community. For routine forensic identification of ancestry, a small number of genetic markers is needed that can be tested quickly and cheaply [4, 5]. In addition to serving in forensic context, estimation of ancestry has become important in the studies of admixed populations. Several AIM sets have been proposed for estimating the admixture between specific ancestral populations such as the African and European genetic contributions to African American populations, and Native American and African contributions to Latino populations [6, 7, 8]. Ancestry estimation also plays a significant role in guiding criminal investigations [9,10]. For example, in 2004 Madrid commuter train bomb attack, ancestry analysis was carried out to identify the origin of the bombers [11]. Furthermore, many studies are investigating the association between ancestry and certain type of diseases [12, 13]. Thus, genetic ancestry analysis is a vast research area using diverse techniques in numerous applications. Most genetic ancestry inference studies focused on developing methods with the aim of distinguishing main continental populations. Some studies identified even very small number of markers to succesfully distinguish continental populations. However, predicting an individual's sub-continental ancestry is still a huge challenge given a number of closely related sub-

populations within the continent. In this work, we aim to address both the continental level and sub-continental level ancestry estimation problem using small set of markers from a single chromosome. Our main goal therefore is to efficiently perform ancestry estimation in a resource-constrained environment.

1.2 Prior Work and Challenges

The most widely used DNA polymorphism in ancestry analysis is single nucleotide polymorphisms (SNPs). Majority of the studies used SNPs as the ancestry informative markers, since they exhibit substantially different allele frequencies between populations from different geographical regions. Other DNA polymorphisms, such as short tandem repeats (STRs) and mitochondrial sequence variation (mtDNA) [14] are not especially powerful for ancestry inference due to their mutational instability. While very large number of SNPs can provide nearly accurate ancestry information for multiple geographic regions, small but robust sets of SNPs are especially useful [15]. Majority of the studies published SNP panels for distinguishing ancestral origins from several continental regions, e.g., Europe, America, Africa and East Asia [16], or between many globally distributed distant populations [17]. Some also proposed small SNP panels, typically in the dozens to hundreds of SNPs which can estimate continental genetic ancestry very accurately [18]. However, very few studies focused on identifying SNP panels for sub-continental ancestry estimation due to the known difficulties of using small SNP panels in distinguishing individuals from closely related populations [19].

Continental ancestry estimation techniques mostly identified SNP markers by examining large enough contrasts in allele frequencies between the continental populations, usually measured by Fixation index (Fst) [20]. Although, continental groups can be distinguished based on high Fst values for the selected set of SNPs [21, 22], this measure is less informative in separating closely

related populations due to small allele frequency differences between intra-continental sub-populations [23-27]. Apart from Fst based ancestry estimation, techniques based on principal component analysis (PCA) [28-30], like EIGENSTART [28], have widespread applications. These methods represent genetic variations by principal component vectors. However, PCA based techniques cannot perform well in case of the data with very large number of individuals as it becomes computationally demanding to compute the eigenvectors [31]. Also, they are not highly efficient due to the requirement of genotyping very large number of SNPs (thousands to millions) to calculate the principal component vectors. For instance, Li et al [32] used 2318 SNPs to infer continental-level ancestry using a principal component derived method. Besides, unsupervised learning (clustering) methods, such as STRUCTURE [33] have been widely used to estimate population structure and assign individuals to different populations. But, these methods perform poorly while inferring population structures in large datasets, due to the requirement of intensive computational time and resources. Some studies used STRUCTURE to develop small panels of SNPS for analyzing ancestral origins for people from a large number of populations, e.g., 73 populations in [34] and 119 populations in [15]. However, they only showed which populations cluster together, without explicit prediction of the sub-populations for the individuals.

Thus, though significant progress has been made in the use of genomic data for ancestry detection, challenges still remain. Although a panel with a relatively small number of SNPs can produce sufficiently accurate continental-level ancestry classification, sub-continental population detection using small set of marker SNPs is still a big challenge. Not much has been done on identifying sets of ancestry informative SNPs (AISNPs) that can accurately distinguish closely related sub-populations, for instance, those from the same continent. This is a difficult multi-class classification challenge, with only a few attempts at the problem. This problem is also related to the issue of

separating admixture populations [7, 35], and recent approaches that have used GWAS (Genome-Wide Association Studies) data [2, 3, 36]. However, we do not address the problem of admixture in the scope of this work and also, we do not use GWAS datasets.

Another challenge is that of computation, and the ever limited resources available in most labs, where such ancestry estimation may be needed. Thus, given resource constraints, it is important to analyze the performance of ancestry inference techniques using the markers from only one or few chromosomes. This will mean that the required sequencing can focus only on the specified chromosome (s), thus minimizing sequencing cost and computation time.

In this thesis, we address the problems of both continental and sub-continental ancestry identification using small SNP panels, with all SNPs in the panel coming from one single chromosome. For this study, we focus on Chromosome 1, since this is the largest chromosome, and thus might provide the best starting point for our exercise. We used the dataset '1000 Genomes Phase III' [37], which contains 26 different populations from 5 different continents. Thus, analyzing the DNA information of Chromosome 1, we exploited machine learning techniques and statistical analyses to identify small sets of SNPs for predicting an individual's continental and sub-continental origin. Particularly, we have addressed a number of different ancestry inference problems, including: (1) Multi-class continental classification, (2) Pairwise/Binary classification between sub-populations, (3) Multi-class intra-continental subpopulation classification, (4) Multi-class all population classification, and (5) Two-stage approach for ancestry prediction integrating information from (1) & (3).

1.3 Contributions of the Thesis

The main contributions of this work can be stated as follows:

1. A single chromosome, particularly Chromosome 1, has been analyzed to identify the powerful candidate SNPs for continental and sub-continental level ancestry estimation. That

is, ancestry inference model developed in this work requires the sequencing of only one chromosome, thus saving time and sequencing cost.

2. We focused on selecting small set of SNP markers for ancestry estimation. Therefore, we have performed pruning of SNP features in multiple initial stages, namely parameter based selection, and outlier based selection, prior to the final selection stage. Two different algorithms have been proposed for final stage of SNP selection. One is 'Correlation-based SNP selection' and the other is 'Random subspace projection based SNP selection'.
3. We used 'Neural network with softmax activation' [64] as the learning algorithm in classification stage of both selection methods.
4. SNPs identified using correlation-based selection approach performs very well in continental-level classification as well as binary classification between closely related sub-populations. In a number of cases of binary classification between sub-populations we can achieve 100% classification accuracy, such as American sub-populations Puerto Rico vs. Peru, African sub-populations Gambian vs. Luhya. But, also there are several challenging cases where binary classification rate is in the range of 60%-70%, such as, British vs. Spanish in Europe.
5. Random subspace projection based approach identifies SNPs that perform well in continental classification as well as sub-continental classification with multiple classes. While performing within-continent multi-class classification of subpopulations, we achieved sufficiently good classification rate using less than 2000 SNPs. For instance, multi-class classification accuracy between seven closely related African sub-populations is 87.6% using 1500 SNPs. Besides, while distinguishing four American sub-populations we achieved 87.5% accuracy. But, distinguishing the sub-populations in South Asia was relatively

difficult. Applying this approach of SNP selection, we have developed a two-layer model for ancestry prediction, which first detects an unknown person's continental origin and then based on the detected continent, it predicts the sub-continental origin from the closely associated sub-populations.

1.4 Organization of the Thesis

The whole thesis is divided into five chapters. Chapter 1 introduces the domain of genetic ancestry inference with a brief discussion of the previous works. The main contributions of this thesis are mentioned in this chapter. In Chapter 2, we provide a detailed review of the existing literatures and discuss about some limitations of the current methods. In Chapter 3, we propose a correlation based SNP selection approach for ancestry prediction. Here, we explain all the pre-processing stages of SNP pruning and the final selection stage based on pairwise correlation of SNPs. The performance of this approach has been evaluated for continental level classification and binary classification of sub-populations. Here, we mention the binary classification accuracies of all sub-population pairs within a continent. In Chapter 4, we propose another approach of SNP selection based on random subspace projection. We applied this approach on the SNPs set obtained after executing the pre-processing and pruning stages mentioned in Chapter 3. The performance of this method has been evaluated for continental level ancestry classification and multi-class classification of closely associated sub-populations. Also, a two-layer ancestry prediction model has been developed for the identification of ancestral origin of unknown individuals using the proposed random subspace projection method. We have evaluated the performance of this two-layer model on the test set of the database used, where the test subjects are from 26 different populations. Finally, we conclude and mention several possible future works in Chapter 5. The overall process of ancestry informative SNP selection proposed in this thesis is depicted in Figure 1-1, which demonstrates how the initial set of

20.1 million SNPs from chromosome 1 have been pruned in several data pre-processing stages (e.g., data cleaning, similarity SNP set removal), and initial pruning stages (parameter based selection and outlier based selection) until they are brought down to a much smaller set of 6404 SNPs. Both correlation based selection approach and random subspace projection based selection approach distinctly identify the continental level and sub-continental level ancestry informative SNPs from the set of 6404 SNPs.

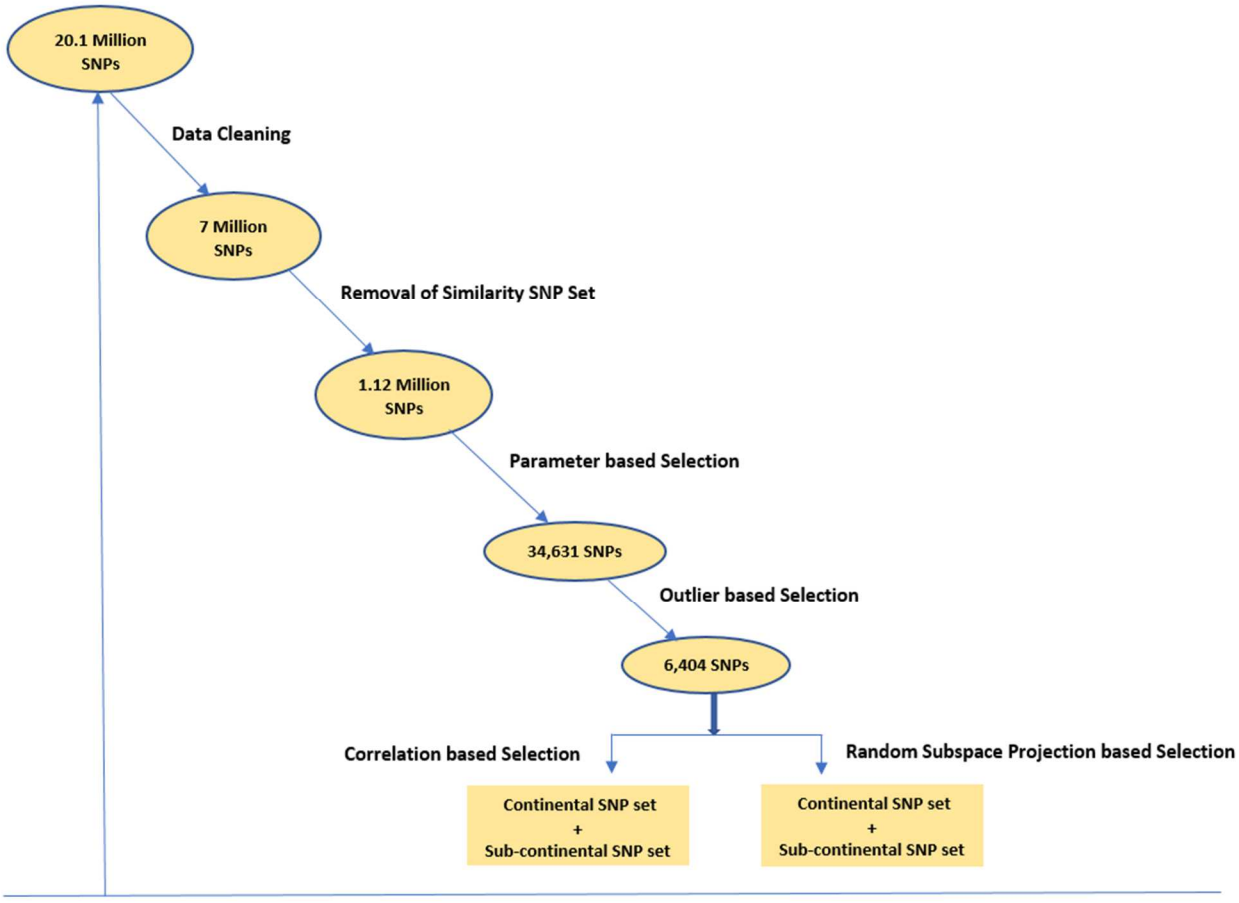


Figure 1-1: Graphical representation of the overall process of SNP selection in multiple stages

Chapter 2: Literature Review

Here, we broadly discuss existing methods in population genetics for detecting individual genetic ancestry. Different ancestry informative markers are also introduced along with their applications in ancestry inference.

2.1 Ancestry informative Markers

Most of our DNA is identical to DNA of others. However, there are inherited regions of our DNA that can vary from person to person. Variations in DNA sequence between individuals are termed as 'polymorphisms'. Human DNA contains different forms of polymorphisms, such as, single nucleotide polymorphism (SNP) [38], short tandem repeat (STR) polymorphism [39], Mitochondrial DNA (mtDNA) polymorphism [40]. Such polymorphisms in human genome can play significant role in genetic ancestry estimation. Ancestry-informative markers represent the polymorphisms for a particular DNA sequence that appear in substantially different frequencies between populations from different geographical regions of the world.

2.1.1 Single Nucleotide Polymorphism

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation among people. A SNP variation occurs when a single nucleotide such as, adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome differs between members of a species. Due to their high abundance in the genome, SNPs serve as the predominant marker type. SNPs have long been used as ancestry informative markers due to containing significant variations in allele frequencies between populations from multiple geographical regions. An individual's genotypes at a group of single-nucleotide polymorphisms (SNPs) can be used to predict that individual's ethnicity or ancestry.

Shriver et al. [41] analyzed 11,555 single nucleotide polymorphisms in 203 individuals from 12 diverse human populations to investigate population stratification. Besides, Seldin's group [42] identified a set of 128 SNPs for identification of the continental origin of people and in estimating the admixture proportions of these individuals. Lins et al. [43] also involved SNP markers to address the admixture problem. They proposed 28 ancestry informative SNPs to infer the genetic admixture in an urban sample of the five Brazilian geopolitical regions.

2.1.2 Short Tandem Repeat

A short tandem repeat (STR), alternatively known as microsatellite, occurs in DNA when a pattern of two or more nucleotides are repeated and the repeated sequences are directly adjacent to each other. The pattern ranges from 2 to 16 base pairs in length. A STR polymorphism occurs when STR loci differ in the number of repeats between individuals. Various studies used STR markers for estimating genetic ancestry. For example, Hashiyada et al. [44] studied distribution of allele frequencies at 17 STRs in 526 unrelated Japanese individuals. Besides, Graydon et al. [45] identified 15 autosomal STR loci to distinguish between Han Chinese, Japanese, Korean, American Caucasian, and South Asian aboriginal group Lodha. Londin et al. [46] also introduced a novel panel of 36 microsatellite AIMs that determines continental admixture proportions. However, STR markers have not become very popular in ancestry inference due to their mutational instability [47].

2.1.3 MtDNA and Y Chromosome Markers

Mitochondrial DNA (mtDNA) and Y chromosomal DNA are uniparental/haploid genetic markers. mtDNA provides information about the female-to-female transmitted lineage, whereas the Y chromosome is informative about male-to-male transmitted lineage. Commercial genetic ancestry testing primarily utilizes these haploid markers to make ancestry inference. Corach et al. used Y

chromosomal and mtDNA markers to infer continental ancestry of Argentineans [48]. In another study [49], a Bayesian approach was applied to infer the ancestry of AfroColombians using mtDNA haplotypes. But, these haploid markers provide much reduced information on individual ancestry in comparison to the autosomal markers.

2.2 Ancestry Inference Methods

Studies on genetic ancestry inference deals with a number of different problems in population genetics. Among them, two most widely addressed problems include (1) detection of continental origin of an individual, and (2) assigning individuals to subpopulations within a continent. Various methods have been proposed to solve either one or both of the problems. Here we present a review on the traditional methods along with a number of recent methods on continental and sub-continental ancestry estimation.

2.2.1 Continental Ancestry Inference

To date, several algorithms have been proposed for estimating continental level ancestry. Bayesian estimation methods are the one of the most prevalent techniques in this domain. STRUCTURE [33], perhaps is the most widely used Bayesian inference method to infer global ancestry, which was developed by Pritchard et al. in 2000. It is a clustering technique that estimates population structure and assigns individuals into population membership groups. Inferring population structures in larger datasets with this method is computationally challenging because it requires intensive computational time and resources. Besides, Phillips et al. [14] proposed a Bayesian classification algorithm based on maximum likelihood to distinguish between three major continental groups-African, European and East Asian. They developed a single-tube 34-plex SNP assay considering the SNPs with highly contrasting allele frequency distributions between these population groups. The performance of this

approach was evaluated on the genome data from 'CEPH Human Genome Diversity Panel' (CEPH-HGDP), resulting in a very low misclassification rate. Another study conducted STRUCTURE analysis to develop a panel of highly discriminative 41 SNPs to infer ancestry among the seven continental regions, such as, Africa, the Middle East, Europe, Central/South Asia, East Asia, the Americas and Oceania [50]. However, this panel was found to be least informative for Eurasian populations, and selection of additional markers was suggested. Nassir et al. also proposed a Bayesian approach for continental distinction between multiple population groups from Oceana, south Asia, east Asia, sub-Saharan Africa, North and South America and Europe using a small panel of 93 SNPs [16]. They computed intercontinental paired F_{st} values using the Weir and Cockerham algorithm [51].

Apart from the Bayesian estimation techniques, principal component analysis (PCA) based methods have been a standard procedure in ancestry inference for a very long time. EIGENSTRAT [28], the most widely used algorithm based on principal component analysis, was developed by Price et al. in 2006. EIGENSTRAT uses PCA to model ancestry variation among the samples. The continental origin variations in allele frequencies among individuals can be elaborated in a lower dimensional space using the derived eigenvectors to score individuals [30]. However, PCA does not estimate the proportional ancestry origin of each individual. Also, it becomes computationally very demanding to compute the eigenvectors while working with datasets of very large number of individuals compared to markers. Paschou et al. [17] also used principal component analysis to identify a small set of 50 PCA-correlated SNPs that effectively assigns individuals into one of nine different populations from HapMap dataset. In addition, FastPop [32] is a recent PCA derived approach which developed a rapid principal component analysis technique for estimating the proportion of intercontinental ancestry for each unknown individual. It conducted the analysis on three different continental

populations, such as, European, Asian and West African using 505 samples from HapMap database along with an additional 19661 samples of own collection. This technique outperformed most of the other PCA based ancestry inference techniques in terms of its superior estimation speed, however this method requires a comparatively large set of 2318 SNPs for measuring continental ancestry.

Moreover, recently supervised machine learning technique has been applied in one of the studies for ancestry estimation, known as ETHNOPRED [53]. It proposed an ensemble classification scheme based on disjoint decision trees that can predict individual's continental ancestry using an ensemble of 3 decision trees involving only 10 SNPs and with an accuracy of 100%. It performed the analysis on HapMap II dataset that contains three distinct continental populations. Also, this model can handle missing SNP values when it is extended to involve 29 decision trees over 149 SNPs. This supervised ancestry estimation method demonstrating superior performance over the previous Bayesian and PCA based methods, indicates the necessity of further studies on ancestry estimation using supervised learning techniques.

2.2.2 Sub-continental Ancestry Inference

Accurate ancestry inference in closely related populations is one of the most challenging problems in population genetics. The number of studies addressing this particular problem is still insufficient. Recent works on ancestry inference are primarily focusing on developing models for distinguishing closely related populations within a continent or, the admixed populations which have been mixing for several generations. ETHNOPRED [53] also addressed sub-continental ancestry identification problem on HapMap III dataset. They performed pairwise/binary classification between subpopulations from Europe, East Asia and Africa showing very high classification rates. Further, they demonstrated multi-class classification result between the North American populations from

diverse origins. But, ETHNOPRED performed very poorly while distinguishing Chinese in Beijing from Chinese in Denver with an accuracy less than 55%. Although, this approach showed inspiring results in terms of estimating sub-continental origin, there is still space for further improvement through involving more population groups and addressing multinomial classification problem between the closely related subpopulations. Graydon et al. [45] also addressed ancestry estimation problem by performing binary classification between similar populations as well as distinctly different populations. 15 autosomal STRs were used as ancestry informative markers in this study instead of SNPs. This study demonstrated sufficiently good classification rate while distinguishing distinct population pairs, such as American Caucasian vs. Japanese or, Han Chinese vs. Indian Lodha, with ≥ 90% accuracy in most cases. However, they could not achieve average classification accuracy >70% while distinguishing closely related population pairs (e.g., Han Chinese vs. Japanese).

Several other studies addressed sub-continental ancestry inference problem from the perspective of admixed populations. Sankararaman et al. [54] proposed an algorithm called LAMP to infer ancestry in admixed populations using sliding windows of contiguous SNPs. This method achieves very high accuracy rates for admixtures from distant ancestral populations, such as African (YRI) vs. American (CEU). However, they cannot perform well in case of closely related admixed populations, e.g., Japanese (JPT) vs. Han Chinese (CHB). Besides, Yang et al. [55] proposed an ancestry inference technique EILA, which uses quantile regression and k-means classifier to distinguish admixed populations. Similar to LAMP, it has higher classification in the binary classification of distant admixed populations, but performs poorly in case of separating closely related population pairs. In contrast, WINPOP [56] is an efficient dynamic programming algorithm which can achieve high accuracy in distinguishing closely related admixed populations as well as

distant population groups. WINPOP performs binary classification between admixed European populations as well as admixed Asian populations with very high classification rate.

Although recent studies on genetic ancestry inference have made progress in terms of addressing sub-continental identification problems, significant challenges still remain in case of distinguishing closely related populations. This problem should be addressed from the viewpoint of both admixed and non-admixed populations.

Chapter 3: Correlation-based SNP Selection

3.1 Background

Feature selection is an essential step prior to applying a learning algorithm on a given task. The performance of a machine learning algorithm often improves significantly if the redundant and irrelevant features are removed from the data prior to learning. A well-known approach for feature selection in machine learning applications is 'variable filtering' [57, 58]. A filter evaluates features according to a statistic based on the general characteristics of the data. With the choice of a threshold, some variables or features are removed. Different filter approaches exist in the literature, such as, t-statistics, F-statistics, Fisher's discriminant ratio, maximum entropy [59], information-theoretic networks [60], correlation-based filters [61, 62], etc. Among them, correlation-based filtering is a popular feature selection technique which aims to identify a set of good features where individual features are highly uncorrelated with each other. In this way, redundant features are being removed from the analysis. Several studies applied the concept of correlation-based filtering for selecting the relevant features in different applications. For example, Hall et al. [61] proposed a correlation based filtering approach which calculates feature-feature correlation using symmetrical uncertainty and finally selects a set of highly predictive features. This algorithm was applied on different datasets of nominal variables. Besides, Whitley et al. [62] designed a correlation-based filtering algorithm which starts by selecting the two features which are least correlated and selects additional features on the basis of their multiple correlation with those already chosen. This algorithm was applied in molecular modeling application for drug design.

In this work, we have designed a correlation-based feature selection algorithm for identifying the set of best SNPs for continental and subcontinental-level ancestry classification.

3.2 Methods

Here, we take a three-stage approach to select the set of candidate SNPs for ancestry estimation. Initially, we apply parameter-based SNP selection, and later refined the selection by using an unsupervised clustering technique (namely, DBSCAN [63]). In the final selection stage, a correlation based filtering approach is applied where we compute pairwise correlation of SNPs to remove the redundant SNPs from the analysis. We apply correlation based SNP selection to identify the important AISNPs for both continental and sub-continental ancestry classification. Once the relevant SNPs are selected, ancestry classification is performed on the test set using the softmax neural network classification scheme [64]. Our continental classification is a five-class classification problem including the continents Europe, Latin America, Africa, East Asia and South Asia. Within each continent there are several closely related sub-populations and accurately distinguishing them is the challenging part. To address the sub-continental classification problem, we have demonstrated pairwise classification of sub-populations within each continent.

3.2.1 Datasets & Pre-processing

For this work, we used data from 1000 Genomes project Phase III [37]. The dataset contains information on 84.4 million variants (SNPs) from all 23 chromosomes for 2504 individuals, from 26 different sub-populations, from five continents. Table 3-1 provides a summary on the different populations, including the number of samples in each of the 26 sub-populations. We focused on analyzing the variants from Chromosome 1 which is nearly 20.1 million SNPs. After data pre-processing steps (e.g., data cleaning), we identified continental and sub-continental ancestry informative SNPs in several stages. The DNA information for the 20.1 million variants (SNPs) from Chromosome 1 of each of the 2504 subjects resulted in a large dataset of size 61.2 GB. At the

beginning, we extracted data from this large dataset and stored them in several smaller tables to be able to conduct our analysis in a MATLAB environment. For each SNP, we extracted their position/loci number, rsID, reference allele, alternate allele (s), and allele information of all 2504 subjects (each person's allele is dip-loid, containing two nucleotides, from different combinations of the four nucleotide bases (A, C, G, T)). Next, we performed data cleaning operations on the extracted data based on the following criteria:

- The SNP loci which contain more than one reference nucleotides have been removed.
- If an alternate allele nucleotide also exists in the reference allele, corresponding SNP position is excluded from the analysis.
- SNP loci where each of the two nucleotides from all the individuals in the dataset both match with the reference allele's nucleotide are excluded from the analysis.

The above steps resulted in the removal of around 13 million SNPs in the cleaning stage. We then performed further analysis using the remaining SNPs. For the purpose of SNP selection, we removed a person's allele information from a SNP position, if the person's both nucleotides at the given position are the same as the reference allele's nucleotide. Consequently, two different sets of SNPs have been observed in the analysis. In one set, each SNP contains the same allele information among all individuals, although this allele information is different from the reference nucleotide. We call this SNP set the 'Similarity Set'. In contrast, in the other set, allele information is not same across all individuals at a given SNP position. We call this set the 'Dissimilarity Set'. Since, for ancestry identification, we need to distinguish between populations with respect to some attributes, SNPs loci which demonstrate greater variation in DNA information among individuals will lead to better

identification performance. Thus, we have chosen only the 'Dissimilarity Set' of SNPs for further analysis.

Table 3-1: Populations in 1000 Genomes Phase III dataset

Population Code	Population Name	Continent	Sample Size
PUR	Puerto Rican	America	104
CLM	Colombian	America	94
PEL	Peruvian	America	85
MXL	Mexican-American	America	64
GBR	British	Europe	91
FIN	Finnish	Europe	99
IBS	Spanish	Europe	107
CEU	CEPH	Europe	99
TSI	Tuscan	Europe	107
CHS	Southern Han Chinese	East Asia	105
CDX	Dai Chinese	East Asia	93
KHV	Kinh Vietnamese	East Asia	99
CHB	Han Chinese	East Asia	103
JPT	Japanese	East Asia	104
PJL	Punjabi	South Asia	96
BEB	Bengali	South Asia	86
STU	Sri Lankan	South Asia	102
ITU	Indian	South Asia	102
GIH	Gujarati	South Asia	103
ACB	African-Caribbean	Africa	96
GWD	Gambian	Africa	113
ESN	Esan	Africa	99
MSL	Mende	Africa	85
YRI	Yoruba	Africa	108
LWK	Luhya	Africa	99
ASW	African-American SW	Africa	61

3.2.2 SNPs Selection

The overall process of SNP selection is explained below in three stages, each building on the results from the previous stage. The SNPs selected in the initial parameter-based selection stage are propagated to the latter stages, where machine learning and statistical analysis are applied to further improve the results, and to prune the selected SNPs to a much-reduced set.

3.2.2.1 Parameter-based Selection

At the beginning, we aimed to identify important markers for each of the 26 populations from the 'Dissimilarity Set' of SNPs. Consequently, we generated a structure array where each row allocates information from one SNP position containing 26 different fields corresponding to the 26 different populations. Each field associated with one population group contains relevant information regarding that group, such as, number of individuals of that group existing at that SNP position (since we removed individuals from a SNP position based on the similarity of their allele with reference nucleotide) and corresponding allele information of those individuals. Next, we calculated two parameters 'α' and 'β' at each dissimilar SNP position for all 26 populations using the following formulas:

$$\alpha_i = \frac{n_p^i}{n_p}$$

$$\beta_i = \frac{f_p^i}{n_p^i}$$

where, $p = 1,2, \ldots\ldots,26$

n_p^i= No. of individuals of population type p existing at SNP i

n_p= Total no. of individuals of population p in training data

f_p^i= Frequency of occurrence of the allele that appears most in population p at SNP i

For any population p, a SNP position i is considered important if at that position $\alpha_i \times \beta_i = 1$ (i.e., α=1 and β=1). Here, α=1 indicates that all individuals of that population exist at SNP i, since none of them has both nucleotides being the same as the reference nucleotide, while β=1 means those individuals also share the same allele information at SNP i. Thus, based on the values of parameters α and β, we identify the best distinguishing SNPs for each population. After we obtain important SNPs set for each population, we take the union of all the 26 sets. The result is a unique set of 38,532 ancestry informative SNPs. From these 38K SNPs, we further removed the SNPs which contain the same allele information across all individuals from all 26 populations in the training set, since SNPs showing no variations between different population groups are not informative in distinguishing them. At the end of this stage, we have 34,631 ancestry informative SNPs in total, all from Chromosome 1.

3.2.2.2 Outlier-based Selection

To further reduce the number of SNPs, we apply an unsupervised cluster-based approach on the results from Stage 1. In particular, we take a contrarian approach: we group the SNPs using a clustering technique. In doing so, we also indirectly identify those SNPs that could not be grouped comfortably into any particular cluster. These are the outlier SNPs that do not seem to be similar to other SNPs, and thus represent good candidates for use in discriminating between ancestries. We use DBSCAN [63] (Density-based spatial clustering of applications with noise) as the basic clustering technique for further selection of important AISNPs which are reasonably distinct from each other. This is a density based clustering technique which does not require the number of clusters of the data to be pre-specified. Given a set of data points in some space, DBSCAN clustering method groups together points that are closely packed together, marking the points as outliers that lie alone in low-

density regions. In our problem, SNPs that contain similar ancestry information are clustered together, while some SNPs are identified as outliers with seemingly unique ancestry information. These outlier SNPs are considered good candidates for distinguishing among populations.

Here, we apply DBSCAN clustering on the 34K SNPs extracted in the previous stage of selection. The algorithm requires three inputs: data matrix D, radius parameter (ε) and neighborhood density threshold (*MinPts*). Data matrix D has 34K number of rows associated with 34K SNPs and each SNP is considered as an object with l dimensions, where l denotes number of training individuals. Each dimension belongs to the allele information of a training subject represented by a number between 1-16, since four nucleotides {A, C, G, T} generate 16 possible allele symbols {AA, AC, …, TT}. The radius parameter ε is measured as the Euclidean distance between two l-dimensional SNP objects and the neighborhood density threshold *MinPts* defines the minimum number of points required to form a cluster. Algorithm 3-1 (adapted from [65]) shows the pseudo code for DBSCAN clustering.

The choice of the two parameters, ε and *MinPts,* requires careful consideration as they have a significant impact on the output clusters. For this problem, we have determined *MinPts*=2, i.e., at least two SNPs will be able to form a cluster if they are within a certain distance ε. And, the value of ε is chosen empirically. We measured the 26-class classification performance for different values of ε for the 80/20 train-test split of the data. For ε =0.1 we obtained the best classification result. Using DBSCAN clustering technique, we have obtained 2378 clusters and 6404 outliers. These 6404 outlier SNPs constitute our new set of candidate SNPs for ancestry identification.

Algorithm 3-1: DBSCAN Clustering Algorithm (adapted from [65]):

Mark all objects as unvisited

DO

Randomly select an unvisited object *p*;

Mark *p* as visited;

IF the ϵ-neighborhood of *p* has at least *MinPts* objects

Create a new cluster *C*, and add *p* to *C*;

Let *N* be the set of objects in the ϵ-neighborhood of *p*;

FOR each point *p* ' in *N*

IF *p* ' is unvisited

Mark *p* ' as visited;

IF the ϵ-neighborhood of *p* ' has at least *MinPts* points,

Add those points to *N*;

IF *p* ' is not yet a member of any cluster, add *p* ' to *C*;

END FOR

OUTPUT *C*;

ELSE mark *p* as noise;

UNTIL no object is unvisited;

3.2.2.3 Correlation-based Selection

As we obtain the set of 6404 SNPs from the clustering technique, we measure the overall 26-class ancestry prediction performance for each individual SNP marker. That is, we perform ancestry prediction using each of the 6404 SNPs, independent of the other SNPs. Of course, we do not expect to produce very good performance for a single SNP. However, the relative performance of the SNPs is a crucial piece of information for our approach. Consequently, a performance matrix X is generated with m=6404 rows, where each row of the matrix is allocated for one SNP representing a six-dimensional vector,

$$\underline{x}^{(i)} = [x_1^{(i)} \; x_2^{(i)} \; x_3^{(i)} \; x_4^{(i)} \; x_5^{(i)} \; x_6^{(i)}]$$

The first element in the vector contains the accuracy of 26-class classification using SNP *i*. The next five elements of the vector are related to the five continents, where each element denotes the percentage of test individuals correctly predicted from a continent. Classification into 26 populations by each SNP has been conducted for 80%-20% train-test split, with *n* individuals. We have used an allele-context feature to represent each SNP during classification, where each SNP's allele-context feature belongs to three possible values: 0, 1, 2. Here, '0' means both nucleotides from an individual at the given SNP location say *i*, are same as the reference nucleotide; '1' means that one of two nucleotides is different from the reference nucleotide; and '2' means that both nucleotides of that individual are different from the reference nucleotide. Allele-context feature vector and class-label vector are denoted for both train and test sets as follows:

$$\underline{a}_{train}^{(i)} = [a_1^{(i)}\, a_2^{(i)} \ldots\ldots a_l^{(i)}]^T \text{ and } \underline{a}_{test}^{(i)} = [a_1^{(i)}\, a_2^{(i)} \ldots\ldots a_{(n-l)}^{(i)}]^T$$

$$b_{train} = [b_1\, b_2 \ldots\ldots b_l]^T \text{ and } b_{test} = [b_1\, b_2 \ldots\ldots b_{(n-l)}]^T$$

Here, *l*=number of training subjects

n-l=number of test subjects

Thus, for $i = 1,2, \ldots\ldots, m$ number of SNPs, the overall performance matrix is represented as,

$$X = [\underline{x}^{(1)}\ \underline{x}^{(2)} \ldots\ldots \underline{x}^{(6404)}]^T$$

Once the performance matrix *X* is created, we calculate the pairwise correlation of SNPs using the associated performance vectors. For example, correlation of SNP *i* and SNP *k* is calculated using the Pearson's correlation coefficient as follows:

$$C = \frac{\sum_{j=1}^{5} (x_j^{(i)} - \bar{x}^{(i)})\,(x_j^{(k)} - \bar{x}^{(k)})}{\sqrt{\sum_{j=1}^{5} (x_j^{(i)} - \bar{x}^{(i)})^2}\ \sqrt{\sum_{j=1}^{5} (x_j^{(k)} - \bar{x}^{(k)})^2}}$$

Here, $x_j^{(i)}$=element of the vector $\underline{x}^{(i)}$ for continent j ($j = 1,2,\ldots,5$)

$\bar{x}^{(i)}$=average of the five $x_j^{(i)}$elements of vector $\underline{x}^{(i)}$

Now, if the correlation coefficient C between SNP i and SNP k is above a certain threshold th, that is, they are highly correlated, one of them is kept in the analysis and the other one is removed. Here, the SNP that provides better classification accuracy in the performance matrix (represented by the first element of vector $\underline{x}^{(i)}$) is considered as "non-redundant", while the other SNP is assumed redundant. The proposed correlation method of SNPs selection is explained below in Algorithm 3-2, using a pseudo code.

Algorithm 3-2: Correlation-based SNP Selection

```
FLAG each SNP as non-Redundant
FOR i = 1 to total number of SNPs
    IF SNP(i) is non-Redundant
        FOR k = i+1 to total number of SNPs
            IF SNP(k) is non-Redundant
                Calculate correlation coefficient C between performance
                feature vectors of SNPs i and k
                IF C > threshold,
                    FLAG SNP(k) as Redundant
                END IF
            END IF
        END FOR
    END IF
END FOR
Remove Redundant SNPs
```

Having described the general procedure for selecting the SNPs, the final step will be to select those that are suitable for continental-level classification, and those that are suitable for more localized discrimination between sub-populations, say from the same continent.

3.2.2.3.1 SNPs Selection for Continental-level Classification

To find the best candidate SNPs for continental level classification, the proposed correlation based SNP selection has been exploited. First, the 6404 SNPs are ranked from highest to lowest based on their classification accuracy in the performance matrix X and 6404×6 performance matrix is rearranged accordingly. Following this rank of the SNPs, we create the order of the SNPs for the initial 'non-Redundant SNP set' in the algorithm and the algorithm is initialized with the best performing SNP. For a certain correlation threshold th, the algorithm is executed to identify the final set of non-Redundant SNPs from the 6404 SNPs. These candidate SNPs represented by the allele-context feature are subsequently used to perform the five-continent classification for 80/20 train-test split. We carried out empirical experiments for a range of values of correlation thresholds and the threshold which provides the best classification performance with the smallest set of SNPs has been finally selected.

3.2.2.3.2 SNPs Selection for Subcontinental-level Classification

When an individual's continental ancestry is known and the individual belongs to any of the two possible closely related sub-populations within that continent, the objective is to identify the accurate sub-population ancestry. In this work, we have selected candidate SNP sets for all possible pairwise classification of sub-populations within a continent exploiting the same correlation algorithm as used in the continental-level ancestry identification. Assume two sub-populations S_1 and S_2 from the same continent j and the goal is to identify a powerful set of candidate SNPs which will be able to

distinguish individuals from these two sub-populations. Now, the 6404 SNPs are ranked from highest to lowest based on the continent j elements ($x_j^{(i)}$) in the performance matrix X and performance matrix is rearranged accordingly. Thus, the correlation algorithm is initialized with the best performing SNP for continent j and for a certain threshold the algorithm is executed to obtain the non-Redundant set of SNPs from the 6404 SNPs. Next, using the allele-context feature of these SNPs, binary classification between the two sub-populations is performed for 80/20 train-test split. Similar to continental-level classification, we tested for a range of values of correlation thresholds and chose the threshold that provides the best classification performance with small set of SNPs.

3.2.2.3.3 Ancestry Classification Algorithm

Having identified the best SNP subsets, a supervised learning algorithm has been applied to perform ancestry classification task. The learning algorithm used in this work for the classification task is 'softmax neural network classifier' [64]. We used this classification scheme for both continental level and sub-continental level classification. In machine learning, softmax regression is a generalization of binary logistic regression that we can use for multi-class classification tasks. In logistic regression, the output labels are assumed to be binary, that is, $y^{(i)} \in \{0,1\}$. The logistic regression hypothesis tries to predict the probability that a given example belongs to the '1' class, i.e., $P(y = 1|x)$ vs. the probability that it belongs to the '0' class, i.e., $P(y = 0|x)$. On the other hand, in softmax regression setting the output label can take K different values, $y^{(i)} \in \{1,2, \ldots, k\}$. Now, the hypothesis estimates the probability for each value of K, i.e., $P(y = k|x)$. Thus, softmax regression is an extension of the logistic regression to the multi-class case. With $K = 2$, softmax regression is same as binary logistic regression. Overall, with softmax regression scheme, we can solve classification problem not just for $K = 2$, but also for many possible values of K.

Softmax regression is often used as activation function in the final layer of a neural network classifier. For a *K*-class classification problem, number of units/nodes in the output layer of the neural network should be *K*. Each of the *K* output nodes gives the probability of a certain class and probabilities from all output nodes sum to 1. In Figure 3-1, we demonstrate a model of a neural network with softmax layer. This particular model shows the case where *K*=3.

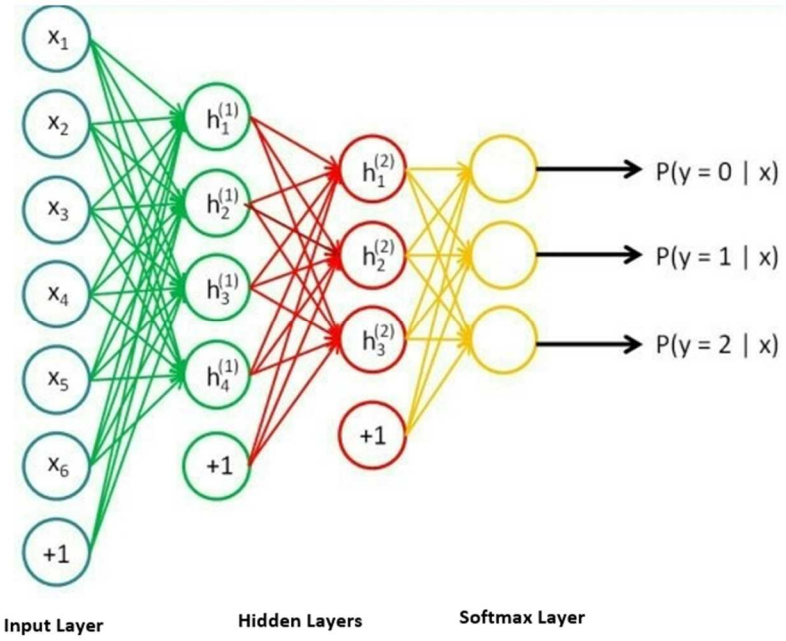


Figure 3-1: Neural network model with softmax activation function

Each output node *i* in final layer of the neural network receives the weighted sum of the inputs from the previous layer with the addition of a bias term, which is denoted as follows,

$$z_i = \sum_j w_{i,j}\, x_j + b_i$$

Where, *j* is the number of nodes in the previous layer. Now to compute the softmax activation at each output node, exponential of the term z_i is calculated for each *i*,

$$t_i = e^{z_i}$$

Finally, activation at output node *i* is obtained by normalizing the exponential term.

$$a_i = \frac{e^{z_i}}{\sum_{i=1}^{K} t_i}$$

Thus, by normalizing the distribution, output from each node *i* falls in the range [0,1]. Here, the class associated with the highest probability value is considered as the predicted output label.

3.3 Experimental Results

We performed experiments using the 1000 Genomes Phase III dataset, with 26 sub-populations, from 5 continents. We evaluated performance of the proposed approach on both continental-level and sub-population-level ancestry prediction/ classification, as described below.

3.3.1 Continental Classification

The five-class classification into five continents -- Europe, America, East Asia, South Asia and Africa has been performed for a range of values of correlation threshold th=0.1 to 0.99 with an interval of 0.01. Figure 3-2 depicts the overall performance in continental-level classification for th=0.4 to 0.99 with 0.01 interval along with the corresponding number of SNPs. The highest performance achieved is 99.19% for th=0.98 with 614 SNPs marked with a red square in the plot. But, since our goal is to rather use a smaller SNP panel for distinguishing continental populations, we searched for the threshold th that provides an optimum performance with less number of SNPs (e.g., 200 or less). From Figure 3-2, we can observe the general trend in performance for the proposed approach. At th=0.4, the system suggests a panel of 10 SNPs, for an overall classification accuracy of about 80%. Performance generally increased with increasing correlation threshold, rising to about 94% accuracy rate, at th=0.82, using 93 SNPs. The best classification result is considered the one for correlation threshold th=0.91, resulting in a classification accuracy of 96.75% with 206

SNPs marked by the magenta square. These 206 SNPs have been considered as our final candidate SNPs for continental-level classification. The confusion matrix for five-class continental classification problem with overall performance of 96.75% is shown in Table 3-2. Our continental classification performance has been compared with other related methods in Table 3-4.

Table 3-2: Confusion matrix for continental ancestry classification

Continents	Europe	America	Africa	East Asia	South Asia
Europe	**94.06%**	3.96%	0.00%	0.00%	1.98%
America	10.94%	**89.06%**	0.00%	0.00%	0.00%
Africa	0.00%	0.00%	**100.00%**	0.00%	0.00%
East Asia	0.00%	0.00%	0.00%	**100.00%**	0.00%
South Asia	1.02%	2.04%	0.00%	0.00%	**96.94%**

3.3.2 Pairwise Classification between Sub-populations

Table 3-3 shows the overall pairwise classification results between sub-populations in each of the five continents in our dataset. The number of SNPs required for each classification has also been noted. From the table, it is evident that in all cases of pairwise classification of closely related populations, we can infer the ethnicity using a small panel of SNPs (less than 200) and for some instances, the accuracy is as high as 100%. For a more detailed analysis, Figure 3-3 (a-f), show the performance of the proposed methods with increasing correlation thresholds, using sub-populations from the continent America. The best performance has been marked with a red square in the figures. As can be observed, it is relatively easy to distinguish between individuals from certain sub-populations, even within the same continent. For instance, Figure 3-3 (a) shows that individuals from Puerto Rico (PUR) can be successfully distinguished from those of Peru (PEL), with an 100% classification accuracy, using only 56 SNPs, under our approach. Also, it is evident from Figure 3-3 (b) that Puerto Ricans (PUR) are easy to distinguish from the Mexican-Americans (MXL), achieving

an accuracy rate of 93.33% with 44 SNPs. Similarly, good pairwise classification results obtained between the populations Columbia (CLM) and Peru (PEL) (Figure 3-3 (d)). It is observed that classification accuracy generally increases with increasing correlation thresholds (and hence more SNPs), but this trend is not monotonic. On the other hand, we can also see some challenging cases, such as Puerto Rico vs. Columbia (Figure 3-3 (c)), where the highest accuracy is about 67% using as many as 89 SNPs. Difficulty is also observed in the binary classification between Columbia (CLM) and Mexico (MXL) (Figure 3-3 (e)), where the highest accuracy is at ~74%, using 37 SNPs. Even increasing the number of SNPs beyond 37 could not improve the result in this case. We have shown comparative results of binary/pairwise classification of sub-populations with other studies in the literature in Table 3-5. The comparative results show the proposed methods are competitive with the state-of-the-art methods, even when using information from just one chromosome.

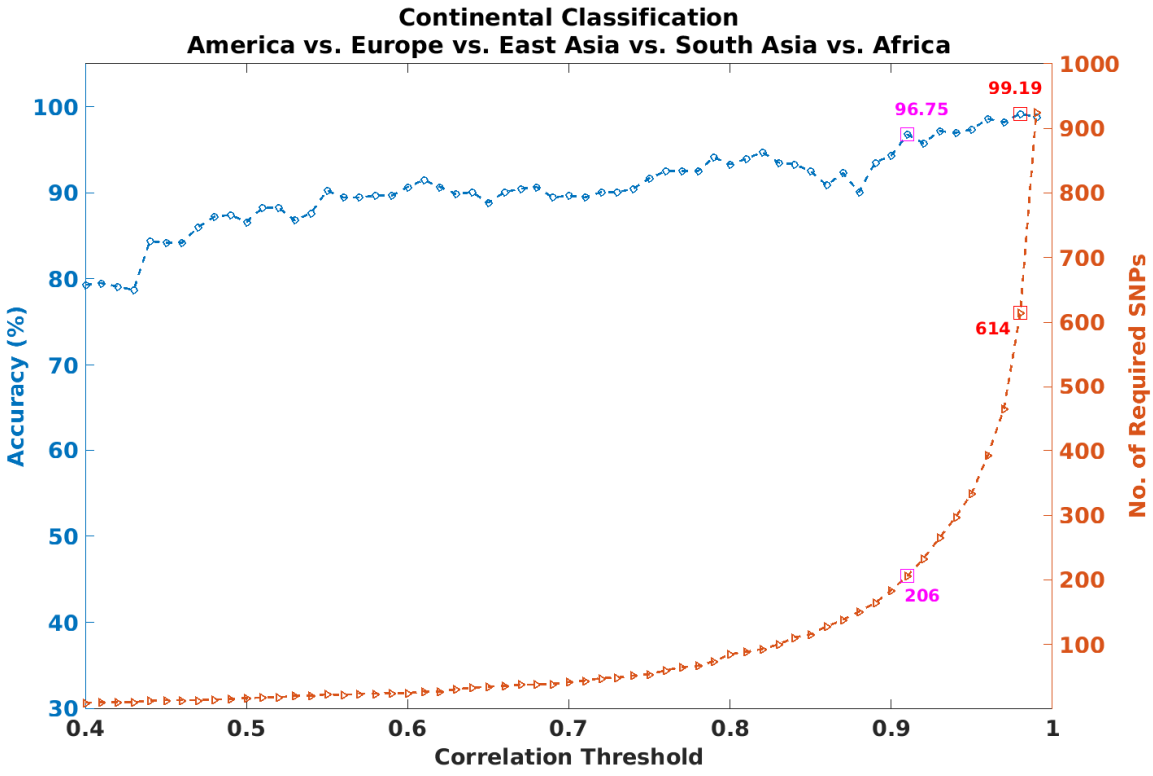
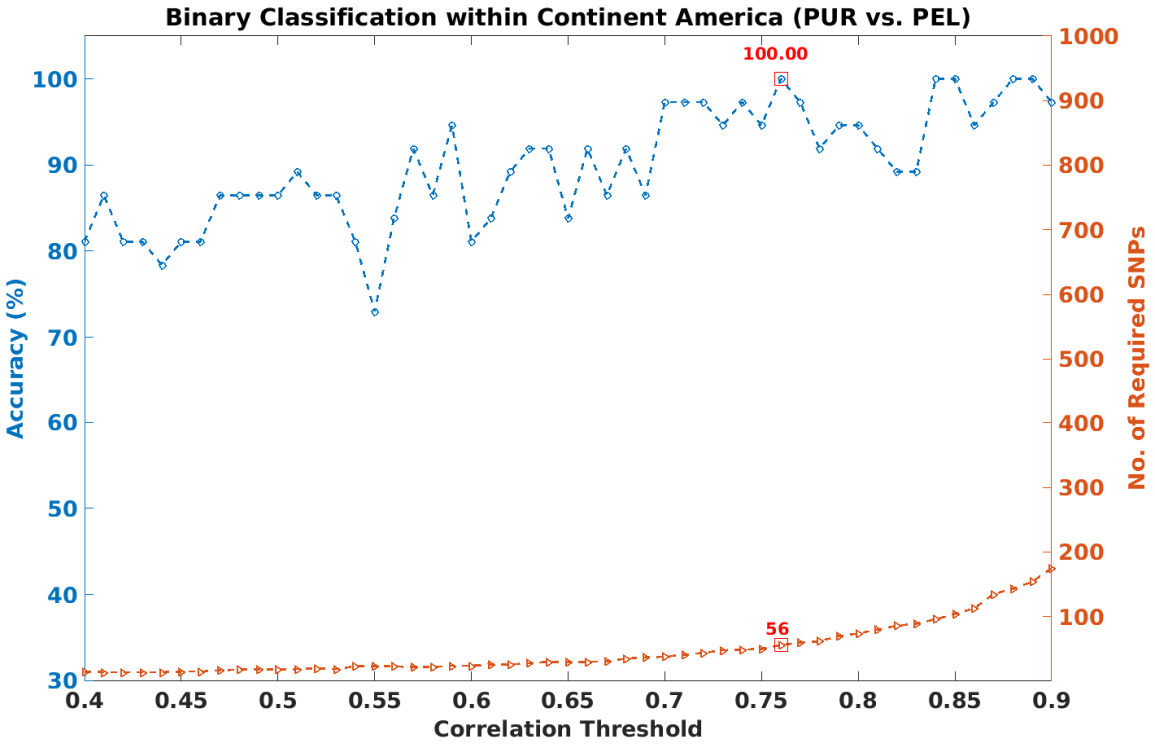


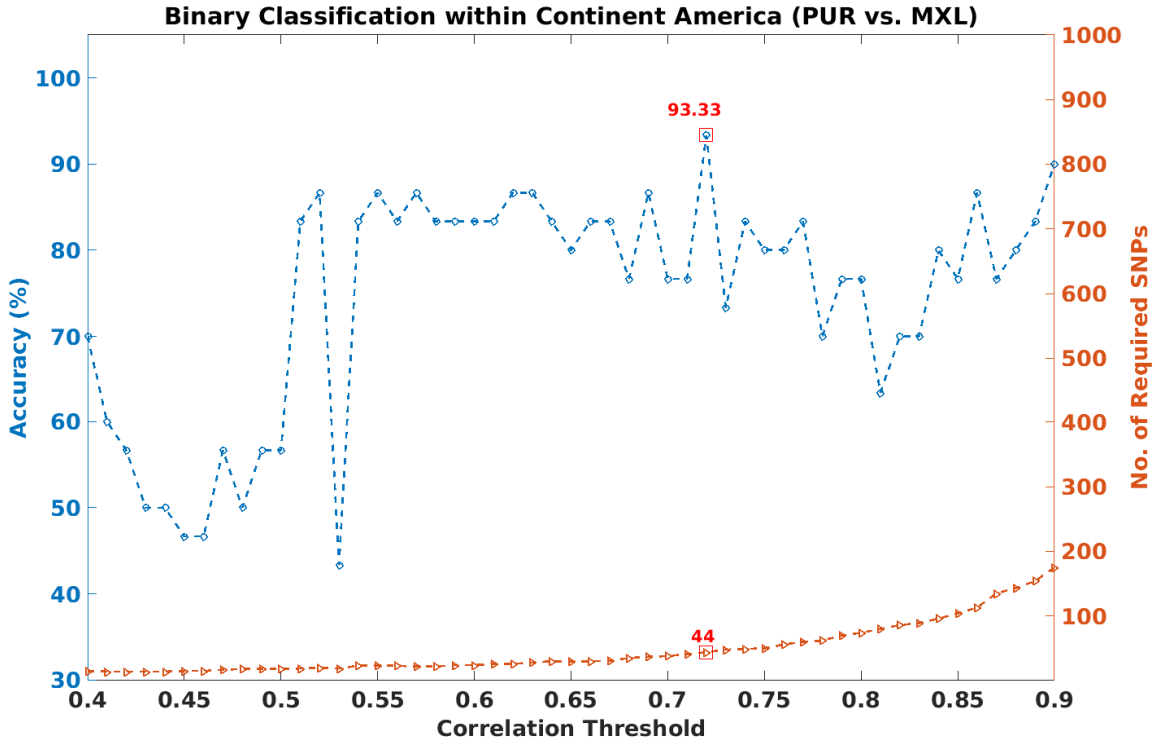
Figure 3-2: Continental classification results with varying thresholds

Table 3-3: Results for pairwise classification between sub-populations

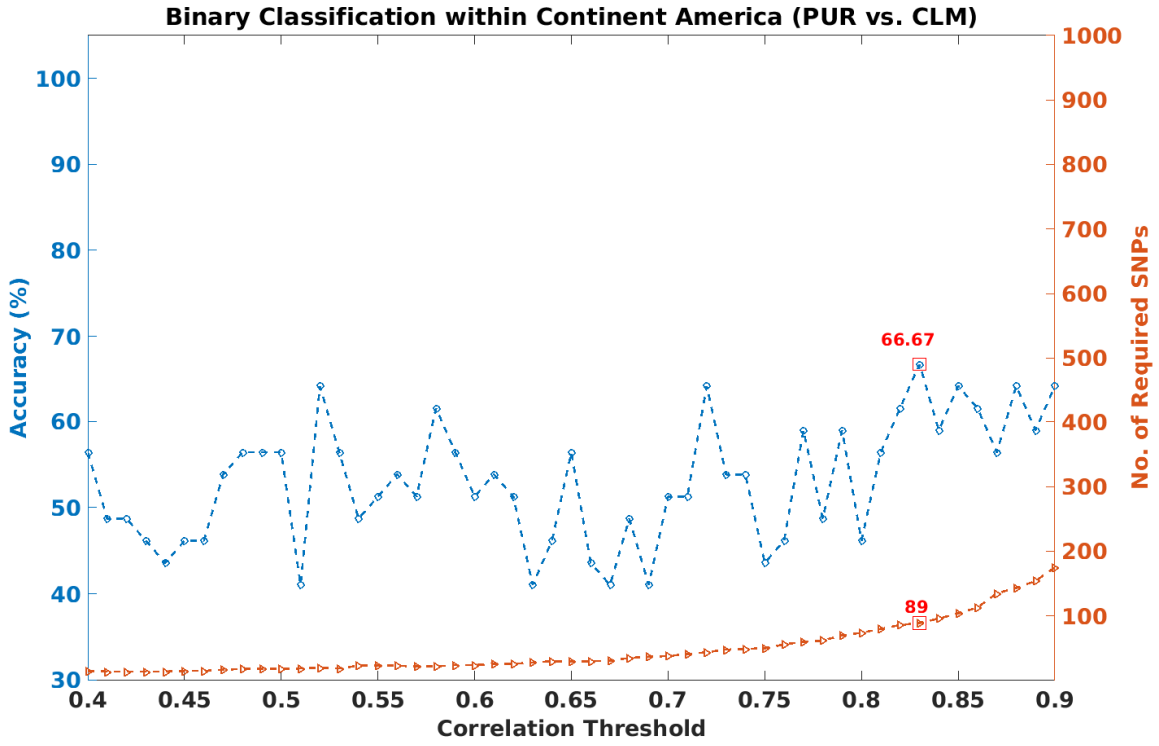
Continent	Sub-populations	Number of SNPs	Correlation Threshold	Accuracy (80-20)
America	PUR-PEL	56	0.76	100.00%
	PUR-MXL	44	0.72	93.33%
	PUR-CLM	89	0.83	66.67%
	CLM-PEL	96	0.84	97.06%
	CLM-MXL	37	0.69	74.07%
	PEL-MXL	96	0.84	84.00%
Europe	GBR-FIN	15	0.47	78.38%
	GBR-IBS	63	0.80	66.67%
	GBR-CEU	30	0.64	67.57%
	GBR-TSI	24	0.61	76.92%
	FIN-IBS	82	0.83	83.33%
	FIN-CEU	130	0.88	80.00%
	FIN-TSI	75	0.82	90.48%
	IBS-CEU	47	0.75	71.43%
	IBS-TSI	82	0.83	77.27%
	CEU-TSI	31	0.67	73.81%
East Asia	CHS-CDX	44	0.73	64.10%
	CHS-KHV	12	0.41	68.29%
	CHS-CHB	30	0.66	64.29%
	CHS-JPT	83	0.84	73.81%
	CDX-KHV	30	0.66	68.42%
	CDX-CHB	120	0.87	76.92%
	CDX-JPT	120	0.87	87.18%
	KHV-CHB	62	0.79	75.61%
	KHV-JPT	92	0.85	82.93%
	CHB-JPT	83	0.84	71.43%
South Asia	PJL-BEB	29	0.65	74.29%
	PJL-STU	57	0.78	62.50%
	PJL-ITU	29	0.65	70.00%
	PJL-GIH	153	0.89	100.00%
	BEB-STU	42	0.72	72.97%
	BEB-ITU	139	0.88	70.27%
	BEB-GIH	113	0.86	100.00%
	STU-ITU	29	0.65	64.29%
	STU-GIH	79	0.82	100.00%
	ITU-GIH	79	0.82	100.00%
Africa	ACB-GWD	47	0.76	76.74%
	ACB-ESN	20	0.56	79.49%
	ACB-MSL	46	0.75	71.43%
	ACB-YRI	43	0.72	80.49%
	ACB-LWK	60	0.79	79.49%
	ACB-ASW	15	0.49	81.48%
	GWD-ESN	46	0.75	77.27%
	GWD-MSL	73	0.82	72.50%
	GWD-YRI	132	0.88	100.00%
	GWD-LWK	132	0.88	100.00%
	GWD-ASW	132	0.88	96.88%
	ESN-MSL	102	0.86	69.44%
	ESL-YRI	132	0.88	100.00%
	ESN-LWK	132	0.88	100.00%
	ESN-ASW	132	0.88	96.43%
	MSL-YRI	38	0.71	100.00%
	MSL-LWK	132	0.88	100.00%
	MSL-ASW	73	0.82	91.67%
	YRI-LWK	28	0.65	78.57%
	YRI-ASW	146	0.89	90.00%
	LWK-ASW	162	0.90	85.71%


Binary Classification within Continent America (PUR vs. PEL)
100.00
56
Accuracy (%)
No. of Required SNPs
Correlation Threshold
100
90
80
70
60
50
40
30
1000
900
800
700
600
500
400
300
200
100
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9

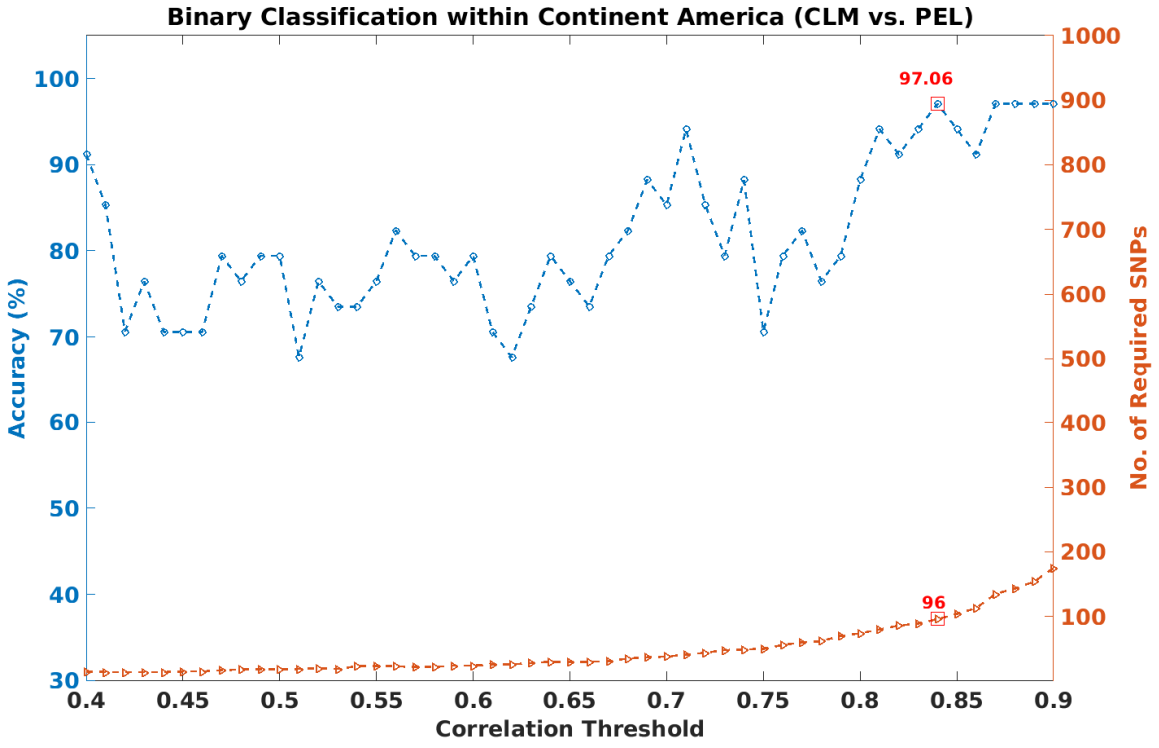
(a)


Binary Classification within Continent America (PUR vs. MXL)
93.33
44
Accuracy (%)
No. of Required SNPs
Correlation Threshold
100
90
80
70
60
50
40
30
1000
900
800
700
600
500
400
300
200
100
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9

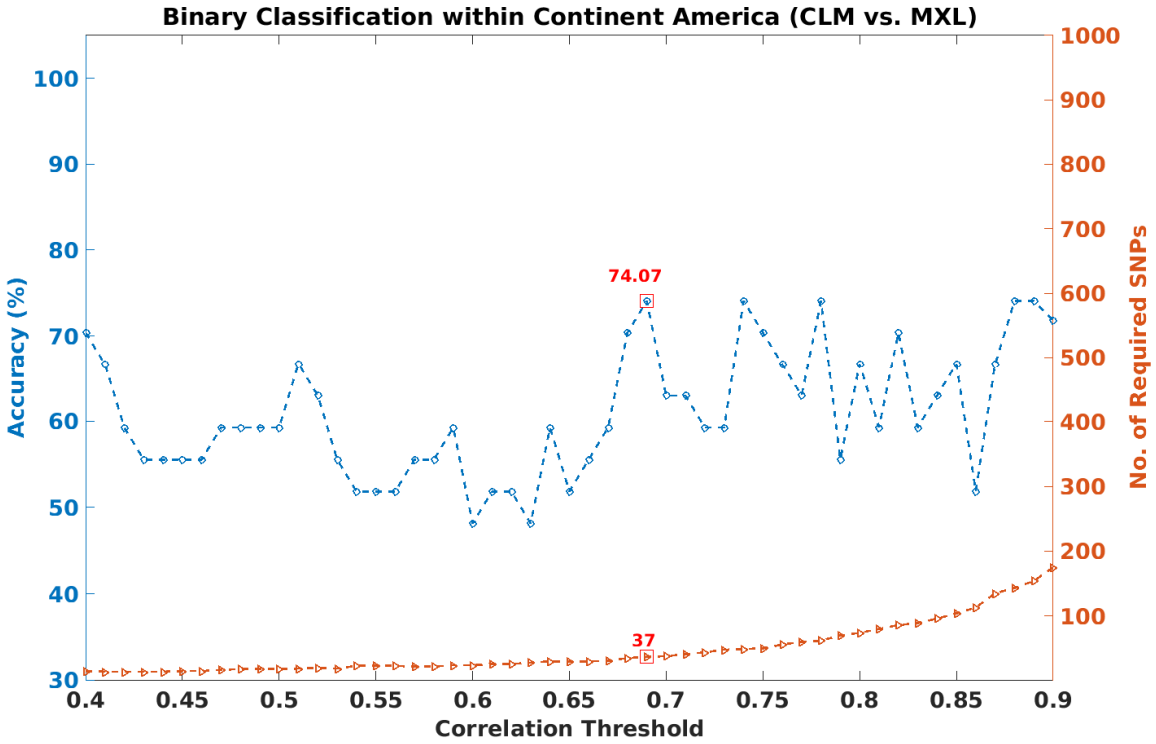
(b)


Binary Classification within Continent America (PUR vs. CLM)
Accuracy (%)
No. of Required SNPs
Correlation Threshold
66.67
89
100
90
80
70
60
50
40
30
1000
900
800
700
600
500
400
300
200
100
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9

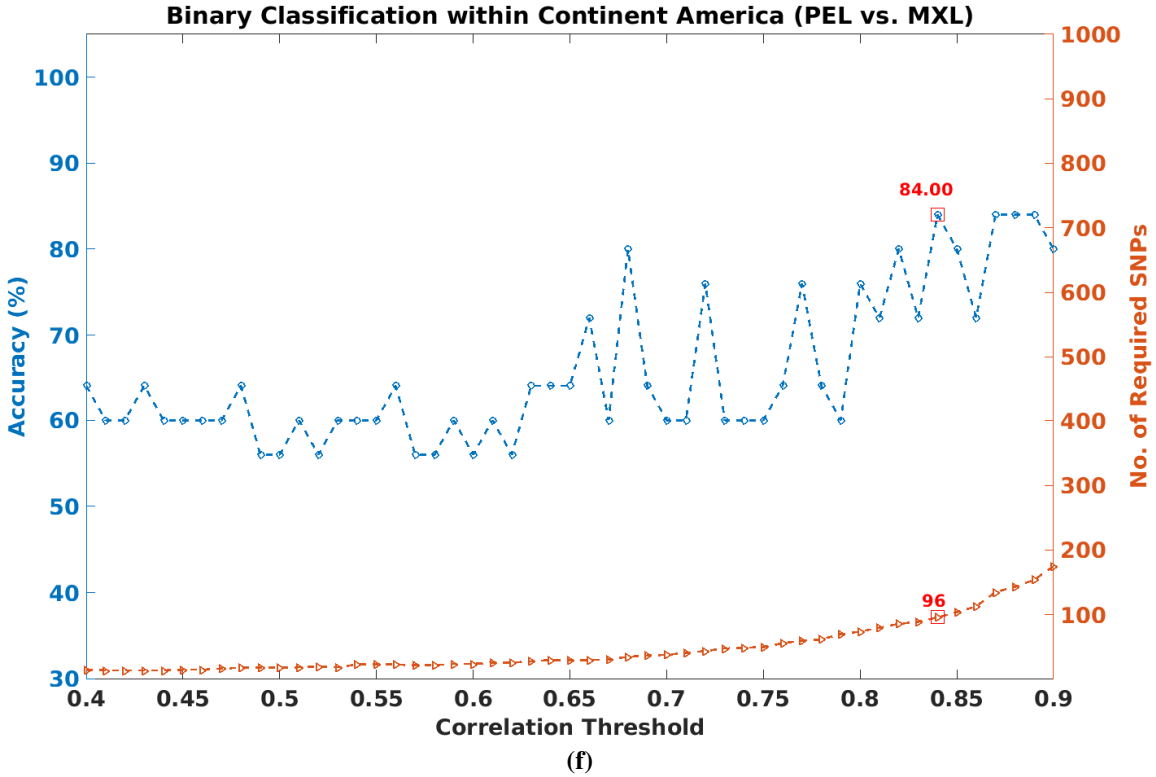
(c)


Binary Classification within Continent America (CLM vs. PEL)
Accuracy (%)
No. of Required SNPs
Correlation Threshold
97.06
96
100
90
80
70
60
50
40
30
1000
900
800
700
600
500
400
300
200
100
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9

(d)



(e)



(f)

Figure 3-3: Pairwise classification results for (a) PUR vs. PEL, (b) PUR vs. MXL, (c) PUR vs. CLM, (d) CLM vs. PEL, (e) CLM vs. MXL, and (f) PEL vs. MXL, with varying thresholds

Table 3-4: Comparative performances in continental ancestry classification (using SNPs)

Basic Method	Data Size	Datasets Used	Classification Rate (%)
[66]	664	Multiple datasets	96.1
[4]	2689	1000 Genome, HGDP, NIST	98.8
[50]	6410	Multiple datasets	81.4
[5]	451	Own Collection	77.0 (+21.6 thresholded out)
Proposed	2504	1000 Genomes Phase III	**99.19 (614 SNPs)**
			96.75 (206 SNPs)

Table 3-5: Comparative performances in pairwise subpopulation classification

Pairwise Sub-populations	Continent	Method	Data size	Datasets	Classification Rate (%)	No. of Attributes
CEU-TSI	EUROPE	[53]	267	HAPMAP III	86.6±2.4	180 SNPs
--	EUROPE	PROPOSED	503	1000 GENOME PHASE III	76.6*	58 SNPs **
CHB-JPT	EAST ASIA	[53]	250	HAPMAP III	95.6± 3.9	877 SNPs
JPT-CHB	EAST ASIA	[45]	9104	OWN COLLECTION	74.9(77.2***)	15 STRs
JPT-KOR	EAST ASIA	[45]	731	OWN COLLECTION	67.9 (63.7***)	15 STRs
CHB-KOR	EAST ASIA	[45]	731	OWN COLLECTION	69.6 (62.4***)	15 STRs
--	EAST ASIA	PROPOSED	504	1000 GENOME PHASE III	73.3*	68 SNPs **
LWK-MKK	AFRICA	[53]	294	HAPMAP III	95.9±1.5	341 SNPs
--	AFRICA	PROPOSED	661	1000 GENOME PHASE III	87.02*	87 SNPs **

*Average accuracy of all pairwise sub-population classifications within the given continent.

**Average number of SNPs required in all pairwise sub-population classifications within the given continent

*** Results obtained without normalization.

3.4 Conclusions

In this chapter, we have developed an ancestry identification system to predict continental origin of an unknown individual and also to distinguish between closely related sub-population pairs within a continent. Here, we are able to construct useful panel of very few SNPs for each case of pairwise sub-continental classification. Later, we conducted an experiment to investigate whether this approach can identify efficient SNPs panel to distinguish multiple closely associated populations. The results obtained from that experiment indicated less effectiveness of this approach for sub-continental classification, when the number of sub-populations is not simply two. Thus, we need to identify better approach for addressing this difficult multinomial sub-population classification problem.

Chapter 4: Random Subspace Projection based SNP Selection

4.1 Background

In the domain of bioinformatics, many studies deal with high dimensional data involving large number of features and limited number of samples, which is popularly known as 'small n large p problem'. For example, microarray datasets measure the gene activity of thousands of genes while the number of samples is limited to several hundred [67]. Due to the high dimensionality of the data and existence of many noisy features, traditional pattern recognition techniques often fail to solve these 'small n large p' problems. Traditional classification algorithms, such as support vector machine (SVM) and the *k*-nearest-neighbor (KNN) classifiers cannot perform well in the presence of increasing number of noisy features, in spite of their ability to handle large number of features. Therefore, various techniques have been proposed to address these problems caused by the high dimensional feature space, including classifier aggregation and feature selection. One of the popular techniques is random subspace method [68], which provides improved classification accuracy by aggregating the power of multiple classifiers. It selects a random subset of features in each pass of the algorithm and constructs a decision tree classifier to predict the unknown samples. The decisions of individual trees are combined to a final decision forest by averaging the estimates of posterior probabilities at the leaves of each tree. Li et al. [69] proposed another technique for high dimensional data classification, where the random subspace idea is exploited to generate the individual classifiers on low dimensional subspaces, and base classifiers are assigned different weights according to their individual performances while aggregating the classifier outputs. Apart from the classifier aggregation techniques, another type of approach in handling high dimensional data is pre-

classification feature selection, which aims to remove the noisy features and selects the features that are discriminative among different classes for the classification analysis. Random KNN [70] is such a feature selection technique, which consists of an ensemble of k-nearest neighbor base classifiers, each constructed from a random subset of the input features. The optimum subset of features is selected through ranking the features using a 'support' measure and further applying a two-stage backward elimination procedure. Another feature selection technique, proposed by Lai et al. [71] also incorporates random subspace selection to identify the finest subset of features, where each base classifier in the reduced subspace provides a weight for all features. The weights obtained from different classifiers are later used to rank the features. In addition, there are many popular gene ranking algorithms which also followed random subspace method, such as, RSM-GR [72] algorithm, where support vector machine (SVM) was used as the base classifier.

In this study, we propose a SNP selection algorithm incorporating the concept of random subspace projection. This is an iterative approach which uses the supervised learning algorithm itself to evaluate the worth of the SNPs. This approach considers the potential interaction among the SNPs in the random subspace. We apply this technique of SNP selection to address 5-class continental classification problem as well as 26-class ancestry classification problem. The 26-class problem is addressed in two separate ways: one-stage approach and two-stage approach.

4.2 Methods

The proposed random subspace projection technique is an iterative SNP selection algorithm, where in each iteration a random subset of SNPs is selected to perform ancestry classification using softmax neural network classifier [64]. The SNP subsets associated with the top classification

performances are chosen and all the SNPs that appeared in those subsets are assigned a rank. Finally, the classifier is evaluated on the test set using the top ranked SNPs in a linearly increasing fashion.

4.2.1 Random Sampling Algorithm for SNP Selection

Genomic datasets typically contain millions of SNPs with limited number of subjects. We have removed many of the noisy SNPs in the preprocessing stages including parameter based selection and outlier based selection (already mentioned in chapter 3) and finally selected 6404 SNPs for further processing. To find an optimum set of ancestry informative SNPs (AISNPs) out of the 6404 SNPs, we apply the proposed iterative random sampling technique. Here, we randomly sample a few number of SNPs, say M from the given set of 6404 SNPs for many number of iterations, for instance N iterations. In each iteration, the randomly selected M SNPs are used to form M-dimensional allele-context feature for each subject t in the dataset, which is denoted as follows:

$$a_t = [a^{(1)}\ a^{(2)} \ldots\ldots\ldots a^{(M)}]$$

Now, for 80%-20% train-test split of the data, the M-dimensional feature space is used to perform multi-class classification exploiting softmax neural network classification scheme. The classification accuracies of all N iterations are stored in a $N\times1$ vector and the corresponding panels of M SNPs are kept in a $N\times M$ matrix. Next, the accuracy elements of the $N\times1$ vector are sorted in a descending order and the rows of the $N\times M$ matrix are rearranged accordingly. After ranking the SNP panels from N iterations, we aim to identify the best contributing SNPs from the top Q SNP panels, where $Q \leq N$. Therefore, we extract the top Q rows of the rearranged $N\times M$ matrix and find the unique SNPs from them.

Let, there are m unique SNPs in top Q rows. Thus, we define a $m\times1$ vector, *count* $= [c^{(1)}\ c^{(2)} \ldots\ldots\ldots c^{(m)}]^T$, where each element denotes the number of occurrence of a SNP in top Q

panels. Next, we rank all the m SNPs based on their values in the count vector. Thus, a SNP is considered powerful for discriminating between populations if it occurs many times in the top Q panels. In this manner, each of the m SNPs is characterized by a ranking. Next, from the sorted m SNPs, we choose the top K SNPs, and utilize the corresponding K-dimensional feature space to perform multi-class classification. Thus, with an increment of K by a certain number δ, classification performance is measured using the top K SNPs until all the unique SNPs are covered. A pseudocode of the overall method is presented in Algorithm 4-1.

Algorithm 4-1: Random Sampling SNP Selection

1. FOR iter = 1 to N
   Take M SNPs randomly
   Extract Allele-context feature for M SNPs
   Measure accuracy
2. END FOR
3. SORT accuracies in descending order
4. Rearrange $N \times M$ SNP matrix based on sorted accuracies
5. Find unique SNPs from Top Q rows of rearranged SNP matrix
6. FOR i=1 to total no. of unique SNPs
   Count (i) = No. of occurrence of i^{th} SNP in $Q \times M$ SNP matrix
7. END FOR
8. Normalized_count = $Count$ / Q
9. Rank each unique SNP by Normalized_count
10. Initialize K to δ
11. WHILE K <= total no. of unique SNPs
    Extract Allele-context feature for Top K SNPs
    Measure Accuracy
    Increment K by δ
12. END WHILE

4.2.2 One-stage Ancestry Classification

The '1000 genomes Phase III dataset' [37] used in this work contains genomes of subjects from 26 different populations in five continents. Predicting the ancestry of an unknown/test individual into one of the 26 populations, without initially detecting the continent of the individual, defines the problem of one-stage 26-class classification. To address this problem, we applied the proposed random sampling algorithm. First, we define two parameters, M and N (say, M=50, N=50000). Next, we execute all the steps of the algorithm till step 9, when we obtain the ranking of each of the m unique SNPs from the top Q panels. In the next step, we initialize the parameter K to δ =100. Then by incrementing K in the interval of δ, the top K SNPs are used to conduct the overall 26-class classification for 80/20 train-test split of the data. We have experimented for several discrete values of Q (Q=100, 1000, 5000, etc.), where $Q <= N$. The best performance for each Q is recorded. Finally, the Q which provides the highest performance, in terms of accuracy and number of SNPs, is considered and the corresponding set of SNPs constitute the best candidate SNPs for one-stage 26-class classification problem.

4.2.3 Two-stage Ancestry Classification

The problem of ancestry classification into 26 populations can also be addressed employing a two-step identification scheme. For an unknown individual, first we detect the continental level ancestry, then the sub-population ancestry is identified by comparing only with the sub-populations within the detected continent.

4.2.3.1 Continental Ancestry Prediction

Since the subjects in our dataset come from five different continents, identifying a person's continental ancestry is a 5-class classification problem. We addressed this problem using the above-mentioned random sampling method. Similar to the one-stage 26-class classification problem, the first 9 steps of the algorithm have been executed until each of the unique SNPs has been assigned a rank. In the next step, parameter K has been initialized to δ=10. The final block of the code iteratively computes the 5-class continent classification accuracy using top K SNPs, with an increment of K by δ, until the value of K becomes equal to the number of unique SNPs in top Q panels. Here, compared to the previous one-stage 26-class problem, in each iteration we considered smaller set of SNPs to measure the continental classification performance, as we know from the existing literature that a few hundreds of SNPs can infer continental ancestry with a very high precision. Thus, for a certain value of Q, the best classification performance is recorded along with the associated number of SNPs. The Q value which provides superior results with respect to classification accuracy and number of SNPs, is chosen and the corresponding set of SNPs are considered as suitable candidates for continental ancestry identification.

4.2.3.2 Sub-population Ancestry Prediction within the Continent

Once the continental ancestry of an unknown subject is identified, the second step of this approach detects the sub-population identity of the individual by conducting multi-class classification within the continent. For example, if an unknown individual is identified as a European, a 5-class classification algorithm is executed to predict that person's sub-continental ancestry out of the five different populations-British, Finnish, Spanish, Italian and CEPH, in the continent Europe. Before executing the second stage of two-step classification, we apply the proposed random sampling algorithm to identify the best set of sub-continental discriminative SNPs, for each of the five

continents. To find the SNPs which are capable of performing sufficiently accurate within-continent sub-population classification, first we execute the steps: 1-9 in the algorithm and set the initial value of the parameter K to δ=100. Next, splitting the subjects from a certain continent (say, America) into train-test set, sub-continental multi-class classification performance is measured over the top K SNPs. Thus, incrementing K by δ iteratively, we continue to measure classification performance in each iteration using the top K SNPs, until the value of K is equal to the total number of unique SNPs in top Q iterations. Here also we have experimented with different values of the parameter Q and selected the result corresponding to the best Q. Thus, for each continent, we identify a set of sub-continental ancestry informative SNPs and utilize them in the second stage of two-step classification to predict a person's sub-population ancestry with already identified continental origin.

4.3 Experimental Results

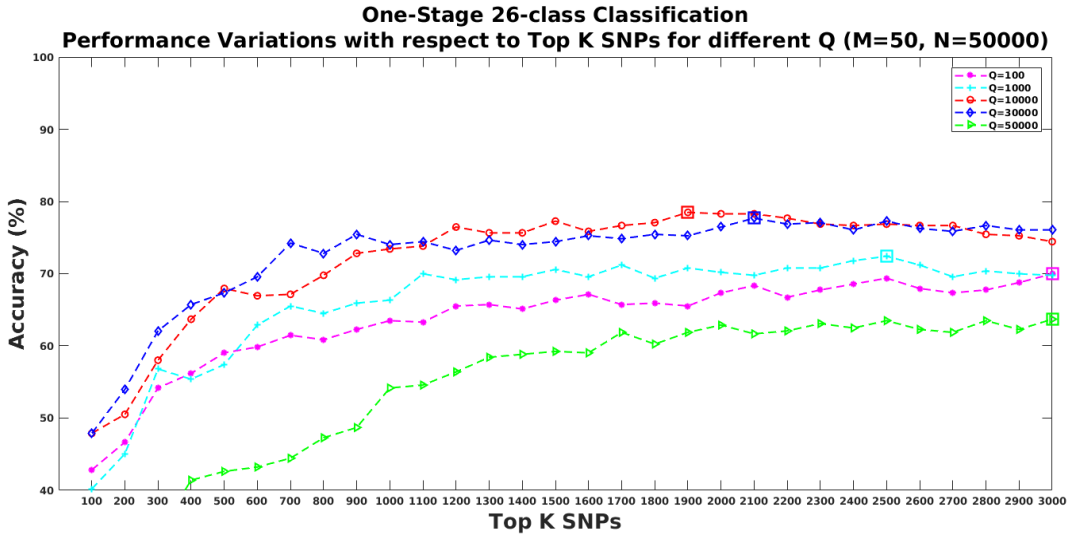
We have evaluated the performance of the proposed random sampling technique for both one-stage and two-stage ancestry classification. All the experiments were performed using the '1000 Genome Phase III' database. The outcomes of different experiments are also explained after careful analysis of the results.

4.3.1 One-stage 26-class Classification

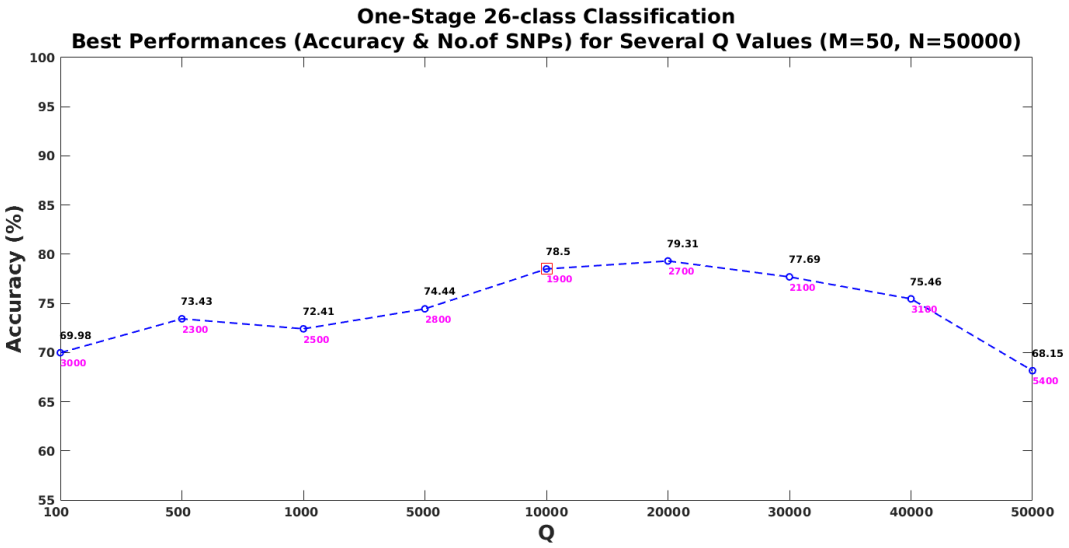
We have demonstrated the results of one-step classification into 26 populations in Figure 4-1(a-b). For our analysis, we have considered M=50 and N=50000. With Q<=N, the values of the parameter Q are chosen over a wide range, with minimum as small as 100 and the maximum equal to 50000. In Figure 4-1(a), the classification performances are depicted for five different values of Q (Q=100, 1000, 10000, 30000 and 50000). For each Q, accuracy is measured on the test set using a certain K

number of top ranked SNPs, with choice of *K* in the interval of 100. From the figure, it is observed that for small value of *Q* (e.g., *Q*=100), the performances over the top *K* SNPs are relatively low, while with increasing value of Q the performances improve. The red curve in the figure demonstrates the best results in one-stage 26-class classification, which corresponds to *Q*=10000. With only 1900 SNPs, classification accuracy of 78.50% is achieved while *Q*=10000. It is also noticed that performances over top *K* SNPs cannot be improved further with higher values of Q (e.g., Q=30000, 50000). With Q=50000 (i.e., Q=N), represented by the green curve, the performances drop to even lower values compared to those for *Q*=100. In Table 4-1, we record the results for each *Q* value in our experiment. For a certain value of *Q*, we mention two types of results: one is the number of unique SNPs available in top *Q* panels of *N* iterations and the classification accuracy achieved using all those SNPs. The other result indicates the best performance over all *Q* values in our experiment, in terms of number of SNPs and corresponding classification accuracy. Results on the best performances for different values of *Q* are also depicted in Figure 4-1(b). Here, we explain the underlying reasons behind the observed trend of the graphs in Figure 4-1(a-b), for fixed *M* and *N* and with varying *Q*. With small *Q*, many SNPs have similar counts of occurrence in top *Q* subsets of *N* iterations, thus they are less likely to be properly ranked. With higher value of *Q*, such as, *Q*=10000, 20000, we observe greater variations between the SNPs in terms of their counts of occurrence in the top panels. This results in a better ranking of the SNPs and better classification results. On the other hand, when *Q* is very large or close to the value of *N* (say, *Q*=50000), SNPs with very high count of the occurrence but occurring mostly in the panels that produced worst results will also achieve higher individual ranking. This eventually deteriorates the overall classification accuracy. The above explanation strongly supports our experimental findings for one-stage 26-class classification problem, as we reach the best classification performance of 78.50% using 1900 SNPs

for Q=10000. Although, it is noticed that in case of one-stage 26-class classification, Q=20000 can produce a slightly higher classification accuracy of 79.31%, but at the cost of much larger number of SNPs. Therefore, for the one-stage classification scheme, we considered the set of 1900 SNPs as the best candidate SNPs capable of classifying samples into one of the 26 populations with accuracy as high as 78.50%.



(a)



(b)

Figure 4-1: (a) One-stage 26-class classification results with varying number of top SNPs, (b) Overall results for one-stage 26-class classification with different choices of parameter Q

Table 4-1: One-stage 26-class classification Results (M=50, N=50000)

	SNPs Coverage (out of 6404 SNPs)		Best Performances	
	No. of Unique SNPs	**Accuracy (80-20)**	**No. of required SNPs**	**Accuracy (80-20)**
Q=100	3378	69.98%	3000	69.98%
Q=500	6237	67.14%	2300	73.43%
Q=1000	6400	67.75%	2500	72.41%
Q=5000	6404	67.34%	2800	74.44%
Q=10000	6404	67.34%	**1900**	**78.50%**
Q=20000	6404	67.34%	2700	79.31%
Q=30000	6404	67.34%	2100	77.69%
Q=40000	6404	67.34%	3100	75.46%
Q=50000	6404	67.34%	5400	68.15%

4.3.2 Two-stage 26-class Classification

The two-stage classification model is built on two successive stages-the second stage is built on the result from the first stage. Prior to developing each stage, we performed experiments with different parameters in the proposed random sampling algorithm. The experimental results associated with both stages of the model are mentioned below.

4.3.2.1 Continental Classification

We propose the approach of 26-class ancestry prediction in two-stages. The approach first identifies an unknown individual's continental ethnicity and next predicts sub-continental origin by classifying the subject into one of the sub-populations within the detected continent. To design such two-step ancestry inference system, initially we identified the best candidate SNPs for continental level classification utilizing the proposed random sampling technique. In Figure 4-2, the continental classification results obtained in this study are graphically represented for fixed values of the parameter *M* and *N*, and different values of the parameter *Q*. For the analysis, *M* and *N* are set to the values of 50 and 50000, respectively. The five curves in the figure correspond to five different values of *Q* (100, 1000, 20000, 40000, 50000), where each curve represents the performances of continental

classification over the top *K* SNPs for a certain *Q*. Similar to one-stage 26-class classification, we observe that for very small or large *Q*, the classification performances are relatively low. The best performances over the top *K* SNPs are achieved for *Q*=20000, indicated by the red curve in the figure. With *Q*=20000, we can perform continental classification with accuracy as high as 97.57% using only 210 SNPs. Accuracy can even increase up to 99.19% using just an additional 170 SNPs. Table 4-2 presents the classification results for each *Q* value in our experiment. Unlike the previous one-step 26-class classification, here we notice that continental classification performances are significantly better with much lower number of SNPs compared to the 26-class classification results. This indicates that multi-class classification with less number of classes is easier to perform while the classes are widely separated, such as continental populations. Thus, we aim to perform continental classification with a small set of SNPs and therefore, consider the 210 SNPs obtained for *Q*=20000 as the best candidates in predicting the continental origin of an unknown/test individual.

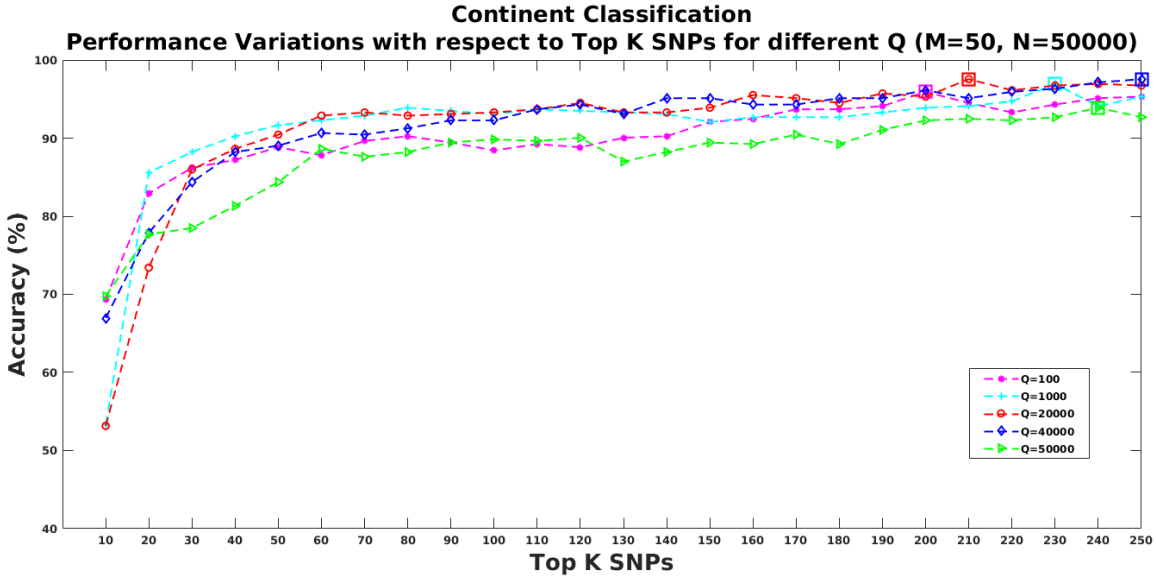


Figure 4-2: Five class continental classification results with varying number of top SNPs

Table 4-2: Continental classification Results (M=50, N=500000)

	SNPs Coverage (out of 6404 SNPs)		Best Performances with Top 250 SNPs		Best Performances with Top 500 SNPs	
	No. of Unique SNPs	Accuracy (80-20)	No. of Required SNPs	Accuracy (80-20)	No. of Required SNPs	Accuracy (80-20)
Q=100	3378	99.80%	200	95.94%	500	98.38%
Q=500	6237	99.59%	250	96.96%	500	98.99%
Q=1000	6400	99.59%	230	96.96%	400	98.38%
Q=5000	6404	99.39%	250	97.57%	480	99.19%
Q=10000	6404	99.39%	250	97.57%	470	98.99%
Q=20000	6404	99.39%	**210**	**97.57%**	**380**	**99.19%**
Q=30000	6404	99.39%	220	97.57%	430	99.19%
Q=40000	6404	99.39%	250	97.57%	420	98.17%
Q=50000	6404	99.39%	240	93.91%	500	96.55%

4.3.2.2 Within Continent Sub-Population Classification

In the two-step ancestry prediction approach, once the continental ancestry is detected, the next step addresses the problem of more localized discrimination between the sub-populations within the detected continent. Using the proposed random sampling technique, we identified powerful sets of sub-continental discriminative SNPs for each of the five continents in our dataset. Each SNP set can make sufficiently accurate prediction regarding the sub-population identity of an unknown individual with known continental origin. In order to obtain such set of subpopulation discriminative SNPs for a certain continent, we first obtain the ranking of the individual SNPs occurred in the top Q subsets of N iterations executing the random selection technique. Then using the softmax neural network classifier, we perform the multinomial subpopulation classification using the top K SNPs, for 80/20 train-test split of the subjects from different sub-populations within a given continent. Figure 4-3 demonstrates the 5-class sub-continental classification (British vs. Finnish vs. Spanish vs. Italian vs. CEPH) performances within continent Europe for five different values of Q (Q=100, 1000, 10000, 30000 and 50000), with M=50 and N=50000. Here, again we notice that the performance curve corresponding to a very small or very large value of Q doesn't indicate the best classification result.

For Q=100 and 50000 (magenta and green curves respectively), the best classification accuracy cannot go beyond 60%. The best classification result is obtained for Q=10000, which is 75.25% using 1400 SNPs, marked by a square on the red curve in Figure 4-3. This result is also evident from Figure 4-4(a), where the best performances for all Q values in our analysis are graphically presented. Thus, these 1400 SNPs are considered as the best informative markers for discriminating between the sub-populations within continent Europe. Therefore, we utilize them to predict an unknown subject's sub-continental origin while the person is initially detected with continental ancestry 'Europe'. Similarly, we performed experiments on other continents to identify the corresponding set of best discriminative SNPs for performing within-continent sub-population classification. Figure 4-4 (b-e) demonstrate the results obtained from four other continents: America, East Asia, South Asia and Africa. Observing the results for all continents, it is evident that in all cases our proposed approach can perform within-continent multi-class classification with sufficiently high accuracy using less than 2000 SNPs. We also list the results from all five continents for different values of Q in Table 4-3.

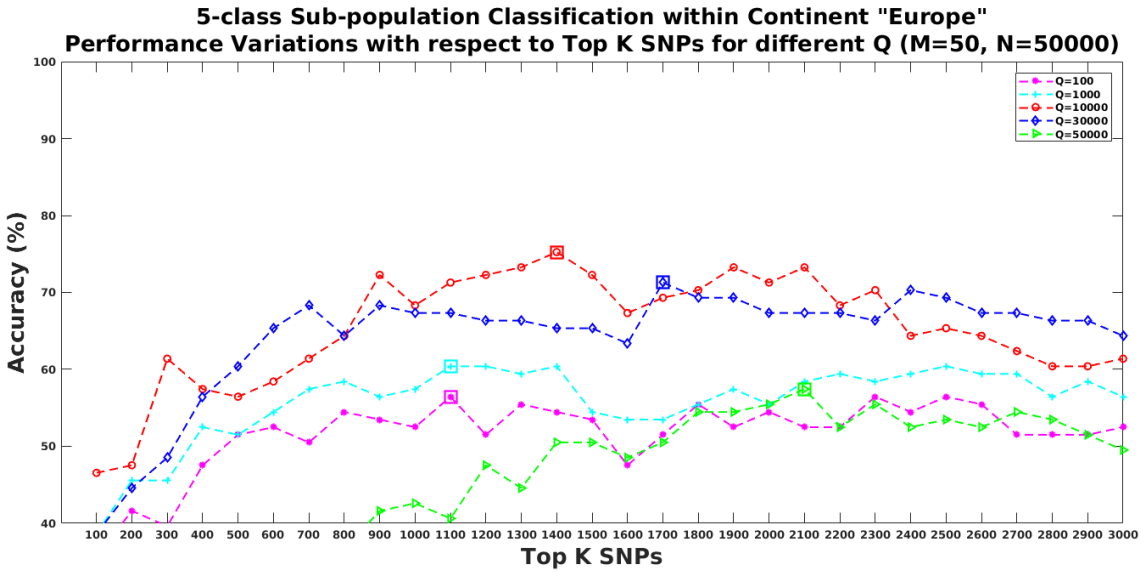
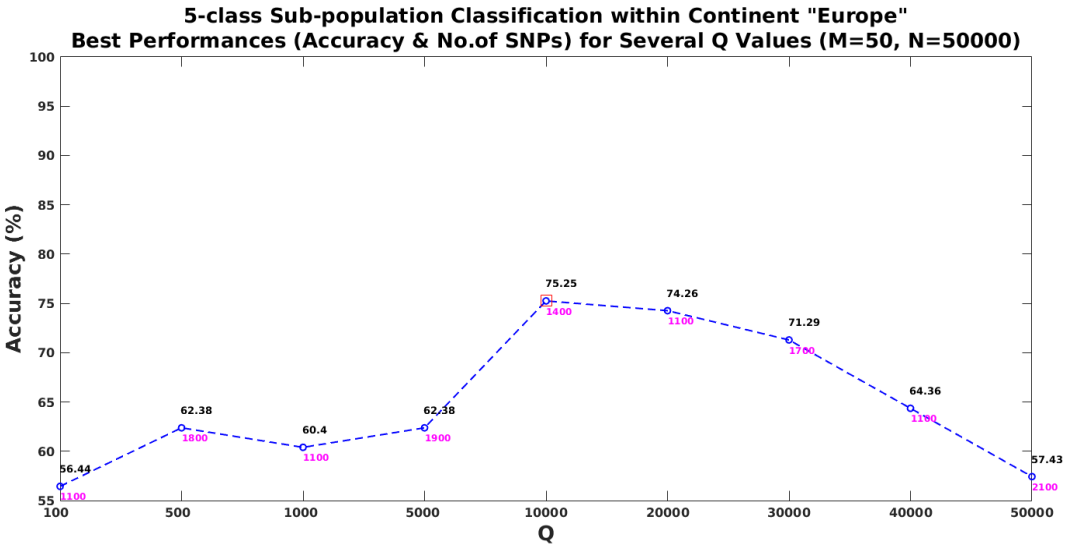
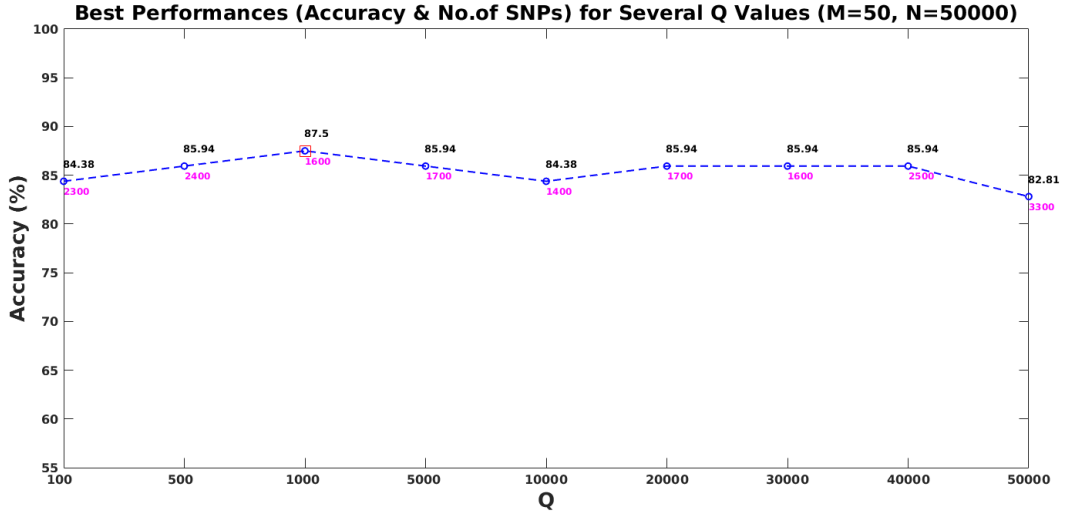


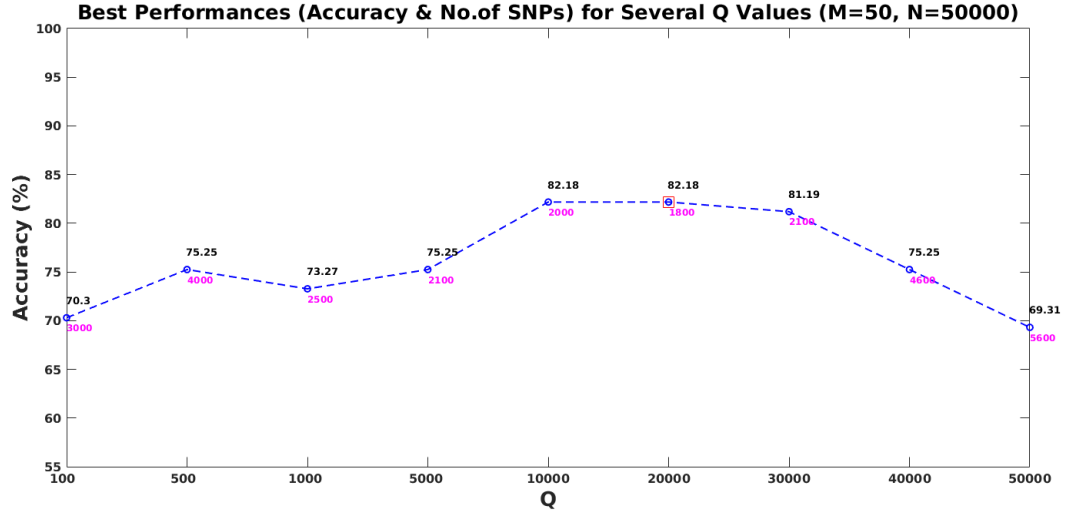
Figure 4-3: Sub-population classification performances within continent 'Europe' with varying number of top SNPs


5-class Sub-population Classification within Continent "Europe"
Best Performances (Accuracy & No.of SNPs) for Several Q Values (M=50, N=50000)
Accuracy (%)
Q
100
95
90
85
80
75
70
65
60
55
100
500
1000
5000
10000
20000
30000
40000
50000
56.44
1100
62.38
1800
60.4
1100
62.38
1900
75.25
1400
74.26
1100
71.29
1700
64.36
1100
57.43
2100

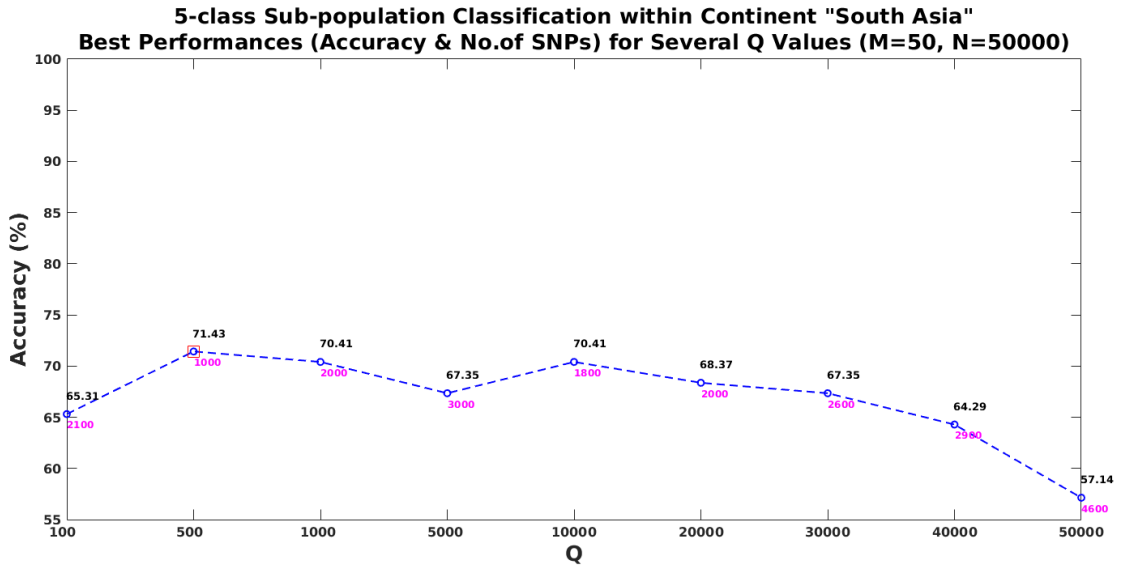
(a)


4-class Sub-population Classification within Continent "America"
Best Performances (Accuracy & No.of SNPs) for Several Q Values (M=50, N=50000)
Accuracy (%)
Q
100
95
90
85
80
75
70
65
60
55
100
500
1000
5000
10000
20000
30000
40000
50000
84.38
2300
85.94
2400
87.5
1600
85.94
1700
84.38
1400
85.94
1700
85.94
1600
85.94
2500
82.81
3300

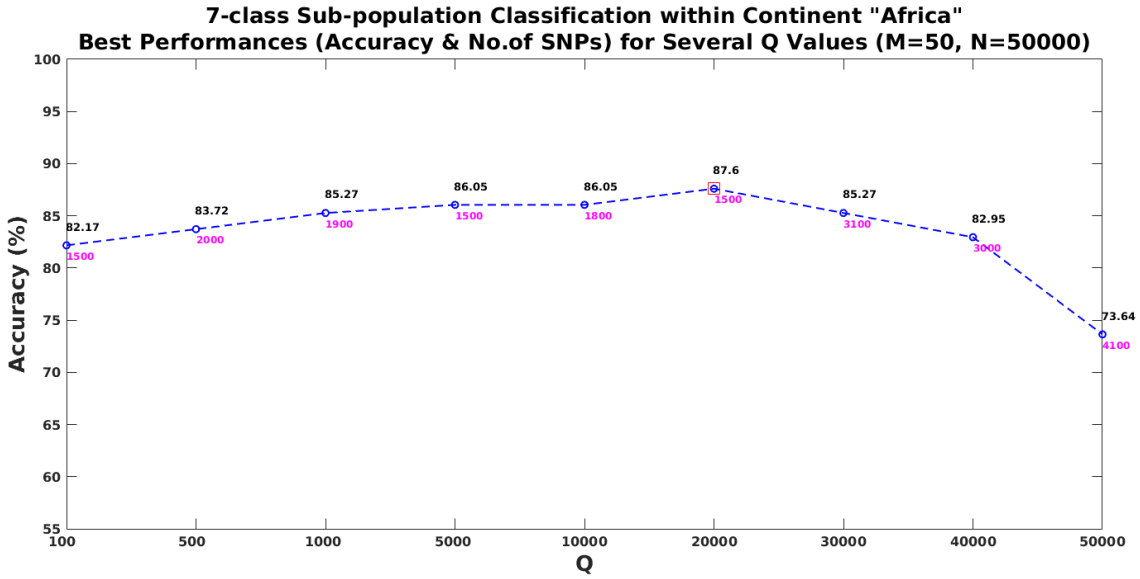
(b)


5-class Sub-population Classification within Continent "East Asia"
Best Performances (Accuracy & No.of SNPs) for Several Q Values (M=50, N=50000)
Accuracy (%)
Q
100
95
90
85
80
75
70
65
60
55
100
500
1000
5000
10000
20000
30000
40000
50000
70.3
3000
75.25
4000
73.27
2500
75.25
2100
82.18
2000
82.18
1800
81.19
2100
75.25
4600
69.31
5600

(c)



(d)



(e)

Figure 4-4: Overall results for sub-population classification within continent (a) Europe (b) America (c) East Asia (d) South Asia (e) Africa, for different choices of parameter Q

Table 4-3: Within continent multi-class sub-population classification results

		SNPs Coverage (out of 6404 SNPs)		Best Performances	
		No. of Unique SNPs	**Accuracy (80-20)**	**No. of Required SNPs**	**Accuracy (80-20)**
Europe	Q=100	3378	52.48%	1100	56.44%
	Q=500	6237	51.49%	1800	62.38%
	Q=1000	6400	49.50%	1100	60.4%
	Q=5000	6404	49.50%	1900	62.38%
	Q=10000	6404	49.50%	**1400**	**75.25%**
	Q=20000	6404	49.50%	1100	74.26%
	Q=30000	6404	49.50%	1700	71.29%
	Q=40000	6404	49.50%	1100	64.36%
	Q=50000	6404	49.50%	2100	57.43%
America	Q=100	3378	85.94%	2300	84.38%
	Q=500	6237	79.69%	2400	85.94%
	Q=1000	6400	78.13%	**1600**	**87.5%**
	Q=5000	6404	78.13%	1700	85.94%
	Q=10000	6404	78.13%	1400	84.38%
	Q=20000	6404	78.13%	1700	85.94%
	Q=30000	6404	78.13%	1600	85.94%
	Q=40000	6404	78.13%	2500	85.94%
	Q=50000	6404	78.13%	3300	82.81%
East Asia	Q=100	3378	69.31%	3000	70.3%
	Q=500	6237	67.33%	4000	75.25%
	Q=1000	6400	67.33%	2500	73.27%
	Q=5000	6404	69.31%	2100	75.25%
	Q=10000	6404	69.31%	2000	82.18%
	Q=20000	6404	69.31%	**1800**	**82.18%**
	Q=30000	6404	69.31%	2100	81.19%
	Q=40000	6404	69.31%	4600	75.25%
	Q=50000	6404	69.31%	5600	69.31%
South Asia	Q=100	3378	65.31%	2100	65.31%
	Q=500	6237	57.14%	**1000**	**71.43%**
	Q=1000	6400	58.16%	2000	70.41%
	Q=5000	6404	58.16%	3000	67.35%
	Q=10000	6404	58.16%	1800	70.41%
	Q=20000	6404	58.16%	2000	68.37%
	Q=30000	6404	58.16%	2600	67.35%
	Q=40000	6404	58.16%	2900	64.29%
	Q=50000	6404	58.16%	4600	57.14%
Africa	Q=100	3378	83.72%	1500	82.17%
	Q=500	6237	79.07%	2000	83.72%
	Q=1000	6400	79.07%	1900	85.27%
	Q=5000	6404	79.07%	1500	86.05%
	Q=10000	6404	79.07%	1800	86.05%
	Q=20000	6404	79.07%	**1500**	**87.6%**
	Q=30000	6404	79.07%	3100	85.27%
	Q=40000	6404	79.07%	3000	82.95%
	Q=50000	6404	79.07%	4100	73.64%

4.3.2.3 Overall Performance of Two-stage Implementation

As we have identified important SNPs for continental classification as well as within-continent sub-population classification, we develop a two-step model for ancestry prediction. For 80/20 train-test split of the dataset, we measure the performance of the proposed two-stage 26-class classification approach on the test set. For an unknown subject, first the continental origin is identified using the 210 continental AISNPs. In the next step, the person is classified into one of the sub-populations within the detected continent by utilizing the sub-continental discriminative SNPs corresponding to that particular continent. For example, if an individual is identified from continent 'Africa', we use a set of 1500 SNPs (identified earlier as strong candidates for discriminating between African sub-populations) to predict that person's sub-continental ancestry. Similarly, if someone is identified from Europe, a set of 1400 SNPs are used to detect the sub-continental origin. The two-step approach is different from the one-step scheme in the way that it doesn't utilize the same set of SNPs to predict every person's ancestry, rather it uses a more specific set of SNPs based on the initial identification of continental ancestry. Overall, 2993 unique SNPs have been used in two-level approach to detect the ancestry of all the test individuals in our dataset. The overall classification accuracy using this two-step ancestry classification scheme is 78.70%.

4.3.3 Comparative Performance Analysis of Two Approaches

We listed the results obtained from one-stage and two-stage classification schemes side by side in Table 4-4. Here, it is noticed that although overall 26-class classification accuracy obtained from two approaches are very close, 78.50% from one-stage approach and 78.70% using two-stage approach, the individual population classification rates and individual continental classification rates are not similar for the two approaches. It is explained earlier that the one-stage approach doesn't include any

initial continental identification, but once we perform the 26-class classification, we can calculate the individual classification rate for each of the continents. We observe that the average continental classification accuracy with one step implementation is 99.60%. On the other hand, in two-stage implementation, we utilize the already identified 210 continental ancestry informative SNPs in the continental identification stage, which can produce average continental classification accuracy of 97.50%. The reason behind higher continental classification rate in one-stage implementation is the use of as many as 1900 SNPs while classifying each individual subject, compared to the use of only 210 SNPs in the continental identification stage of two-stage implementation. With higher number of SNPs one-stage approach produces quite negligible error in continental class identification, but individual population classification rate drops very low for several instances. For example, British (GBR) classification rate in one-stage approach is 41.18% and African-Caribbean (ACB) classification accuracy is 42.11%. Besides another African population ASW suffers from low classification accuracy of 50%. In all three cases, two-stage approach provides better classification performance, about 10% improvement in the first two cases and 25% improvement in the third case. Also, we observe significant performance improvement for the populations CHS, PJL and ITU while using two-stage approach instead of the one-stage approach. However, two-stage scheme performs poorly in case of classifying BEB and PUR populations with accuracy of 62.50% and 76.19% respectively in comparison to the corresponding 81.25% and 95.24% achieved accuracy by one-stage approach. Thus, we can conclude that due to the use different continent-specific set of sub-population discriminative SNPs in the two-stage scheme, this approach performs well for most populations with few exceptions, and doesn't cause any individual population classification accuracy to go below 50% unlike the one-stage scheme. The classification performances of each individual population can be

better visualized from the confusion matrix results shown in Figure 4-5 (a) & (b), for one-stage and two-stage schemes, respectively.

Table 4-4: Comparative Performances for One-stage and Two-stage Implementations

Populations	One-stage Approach	Two-stage Approach
Individual Population Classification Rates		
GBR	41.18%	52.94%
FIN	95.00%	85.00%
IBS	77.27%	72.73%
CEU	60.00%	55.00%
TSI	81.82%	86.36%
PUR	95.24%	76.19%
CLM	94.44%	83.33%
PEL	93.75%	93.75%
MXL	66.67%	66.67%
CHS	61.90%	71.43%
CDX	66.67%	66.67%
KHV	80.00%	85.00%
CHB	80.95%	85.71%
JPT	95.24%	100.00%
PJL	57.89%	73.68%
BEB	81.25%	62.50%
STU	57.14%	51.52%
ITU	57.14%	71.43%
GIH	100.00%	90.48%
ACB	42.11%	52.63%
GWD	91.67%	96.10%
ESN	100.00%	95.00%
MSL	93.75%	87.50%
YRI	90.91%	90.91%
LWK	100.00%	100.00%
ASW	50.00%	75.00%
Overall Classification Accuracy	**78.50%**	**78.70%**
Continental Classification Rates		
Europe	100%	96.04%
America	100%	90.63%
East Asia	99.01%	100%
South Asia	100%	97.96%
Africa	100%	100%
Average Continental Accuracy	**99.60%**	**97.57%**

	GBR	FIN	IBS	CEU	TSI	PUR	CLM	PEL	MXL	CHS	CDX	KHV	CHB	JPT	PJL	BEB	STU	ITU	GIH	ACB	GWD	ESN	MSL	YRI	LWK	ASW
GBR	**41.18%**	11.76%	11.76%	23.53%	11.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
FIN	5.00%	**95.00%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
IBS	9.09%	0.00%	**77.27%**	4.55%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CEU	15.00%	5.00%	10.00%	**60.00%**	10.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TSI	0.00%	4.55%	13.64%	0.00%	**81.82%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PUR	0.00%	0.00%	0.00%	0.00%	0.00%	**95.24%**	4.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLM	0.00%	0.00%	0.00%	0.00%	0.00%	5.56%	**94.44%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PEL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**93.75%**	6.25%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MXL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	11.11%	22.22%	**66.67%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CHS	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**61.90%**	4.76%	9.52%	23.81%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CDX	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**66.67%**	27.78%	0.00%	0.00%	0.00%	5.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
KHV	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5.00%	10.00%	**80.00%**	5.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CHB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.52%	0.00%	0.00%	**80.95%**	9.52%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
JPT	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.76%	0.00%	0.00%	0.00%	**95.24%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PJL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**57.89%**	21.05%	5.26%	15.79%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
BEB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**81.25%**	6.25%	12.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
STU	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.76%	14.29%	**57.14%**	23.81%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ITU	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	28.57%	**57.14%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GIH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ACB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**42.11%**	10.53%	21.05%	21.05%	5.26%	0.00%	0.00%
GWD	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**91.67%**	0.00%	8.33%	0.00%	0.00%	0.00%
ESN	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
MSL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.25%	0.00%	**93.75%**	0.00%	0.00%	0.00%
YRI	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.55%	0.00%	0.00%	**90.91%**	4.55%	0.00%
LWK	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%
ASW	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	37.50%	12.50%	0.00%	0.00%	0.00%	0.00%	**50.00%**

(a)

	GBR	FIN	IBS	CEU	TSI	PUR	CLM	PEL	MXL	CHS	CDX	KHV	CHB	JPT	PJL	BEB	STU	ITU	GIH	ACB	GWD	ESN	MSL	YRI	LWK	ASW
GBR	**52.94%**	5.88%	5.88%	23.53%	11.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
FIN	0.00%	**85.00%**	0.00%	5.00%	0.00%	5.00%	0.00%	0.00%	5.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
IBS	4.55%	0.00%	**72.73%**	4.55%	9.09%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CEU	20.00%	5.00%	5.00%	**55.00%**	15.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TSI	0.00%	0.00%	13.64%	0.00%	**86.36%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PUR	0.00%	0.00%	9.52%	0.00%	4.76%	**76.19%**	4.76%	0.00%	4.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLM	0.00%	5.56%	0.00%	0.00%	0.00%	11.11%	**83.33%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PEL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**93.75%**	6.25%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MXL	0.00%	0.00%	0.00%	0.00%	11.11%	0.00%	11.11%	0.00%	**66.67%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	11.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CHS	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**71.43%**	4.76%	0.00%	23.81%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CDX	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**66.67%**	33.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
KHV	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	**85.00%**	5.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CHB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	0.00%	0.00%	**85.71%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
JPT	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PJL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**73.68%**	15.79%	5.26%	5.26%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
BEB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**62.50%**	25.00%	12.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
STU	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.76%	9.52%	**51.52%**	34.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ITU	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.76%	4.76%	19.05%	**71.43%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
GIH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.76%	0.00%	4.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**90.48%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ACB	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**52.63%**	5.26%	5.26%	26.32%	5.26%	0.00%	5.26%
GWD	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**96.10%**	0.00%	0.00%	3.90%	0.00%	0.00%
ESN	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5.00%	**95.00%**	0.00%	0.00%	0.00%	0.00%
MSL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.25%	6.25%	**87.50%**	0.00%	0.00%	0.00%
YRI	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.55%	0.00%	0.00%	0.00%	**90.91%**	4.55%	0.00%
LWK	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%
ASW	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**75.00%**

(b)

Figure 4-5: (a) Confusion Matrix for one-stage 26-class classification & (b) Confusion Matrix for two-stage 26-class classification

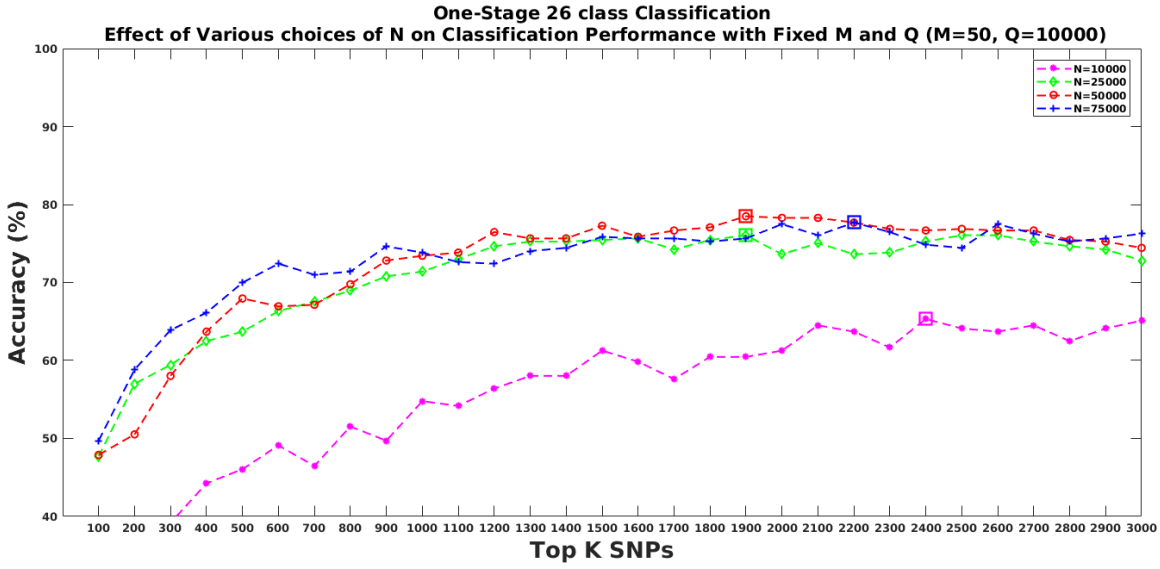
4.3.4 Choice of the Parameters *M* & *N*

In our analysis, we have demonstrated all the results regarding multi-class ancestry identification (one-step /two-step 26-class classification and continental classification) for a certain value of the parameters, *M* and *N* (*M*=50 and *N*=50000). However, we have conducted empirical experiments with several other choices for *M* and *N*, and finally considered those values which provide the

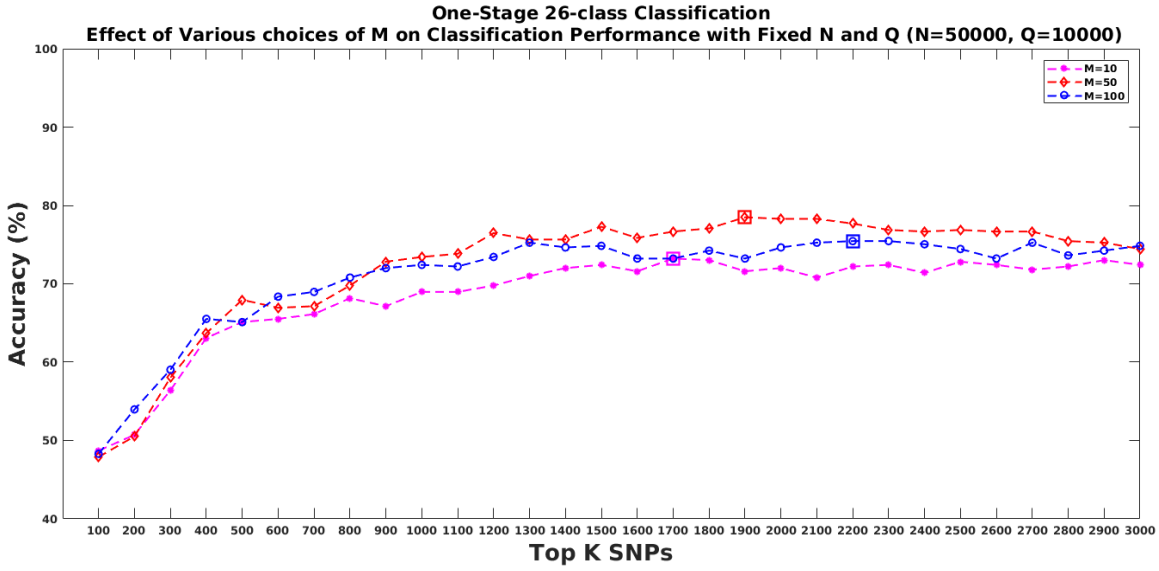
optimum performance in most circumstances. The following Figure 4-6 (a & b) exhibit the graphical evidences to support our choices of *M* and *N*. In Figure 4-6 (a), we have shown how classification performances over top *K* SNPs vary with different choices of the parameter *N* in the case of one-stage 26-class classification. Here, parameter *M* has been kept fixed at 50 and parameter *Q* has been set to 10000 (*Q*<=*N*). Classification performances have been measured against top *K* SNPs for four different values of *N* (10000, 25000, 50000 and 75000). From the figure, it is observed that for *N*=10000, represented by the magenta curve, the performances are quite low. As we consider a higher value for *N*, performance curve goes up in the vertical axis. The curve corresponding to *N*=50000 (red curve) seems the best performing one with *M*=50 and *Q*=10000. Below, we also provide convincing explanations in favor of such observations in our experiments. When N is as small as 10000, that is the random sampling algorithm is executed for 10000 iterations, it is less likely to obtain all possible combinations of *M* SNPs out of initial 6404 SNPs and hence many contributing panels of *M* SNPs might be absent from the analysis. As a result, the unique SNPs in top *Q* iterations cannot be not properly ranked and therefore produce lower classification performance. Conversely, when the algorithm is run for more iterations (higher *N*), it captures more possible combinations of *M* SNPs, and thus the chance of having higher performing panels of *M* SNPs also increases. With the better contributing panels in top *Q* iterations, SNPs are more likely to be properly ranked and thus yield better results. In Figure 4-6(a), we observe that better classification results are obtained when the value of *N* is changed from 10000 to 25000 and 25000 to 50000. However, it is also noticed that increasing *N* beyond a certain value cannot guarantee further increase in classification performance. In our case, as we set *N* to 75000, we don't observe much improvement over the results from *N*=50000. In fact, with *Q*=10000, the best performance achieved for *N*=50000 is slightly higher over the best performance for *N*=75000. Thus, we have concluded

that running our random sampling algorithm for 50000 iterations along with a suitable choice of *M* can provide sufficiently high multi-class ancestry classification performance.

As we have conducted experiments to identify the best choice for parameter N, likewise we have executed our algorithm for different values of the parameter *M* (*M* < ½*6404) while keeping *N* constant. In Figure 4-6 (b), classification performances of one-step 26-class classification are plotted over the top *K* SNPs for three different choices of *M* (*M*=10, 50 and 100) with *N* fixed at 50000 and *Q* set to 10000. From the figure, it is evident that the performance curve corresponding to either *M*=10 or *M*=100 is lower in height along the vertical axis compared to the curve associated with *M*=50 (red curve). The underlying reason behind such outcome is also explained here. In our random sampling algorithm, if we consider *M* to be very small (say, *M*=10), the quantity '6404 choose *M*' is also small, i.e., the total number of possible combinations of *M* SNPs is small. Thus, it is more likely to obtain same combination of SNPs repeatedly in the top *Q* iterations, which might lead to improper ranking of SNPs. For example, if a SNP provides sufficiently good classification accuracy combining with the same set of other SNPs multiple times in top *Q* iterations, it might not be the best SNP despite its very high occurrence in top Q subsets. On the other hand, when M is as high as 100, in spite of the larger value for '6404 choose *M*', many noisy SNPs are being included with the good SNPs to perform classification in each iteration of the algorithm. This might also cause improper ranking of SNPs and therefore yields lower classification performance. From Figure 4-6 (b), it is evident that with a proper choice of parameter *N*, *M*=50 can produce superior classification results compared to *M*=100, utilizing a further reduced set of SNPs (indicated by the square on the red curve).



(a)



(b)

Figure 4-6: (a) Experimental results for different choices of N in one-step 26-class classification with constant M and Q, (b) Experimental results for different choices of M in one-step 26-class classification with constant N and Q

4.4 Random Sampling vs. Correlation Algorithm

Here we compare between the random subspace based SNP selection algorithm and the correlation based SNP selection algorithm in terms of multi-class ancestry classification performances and computation time.

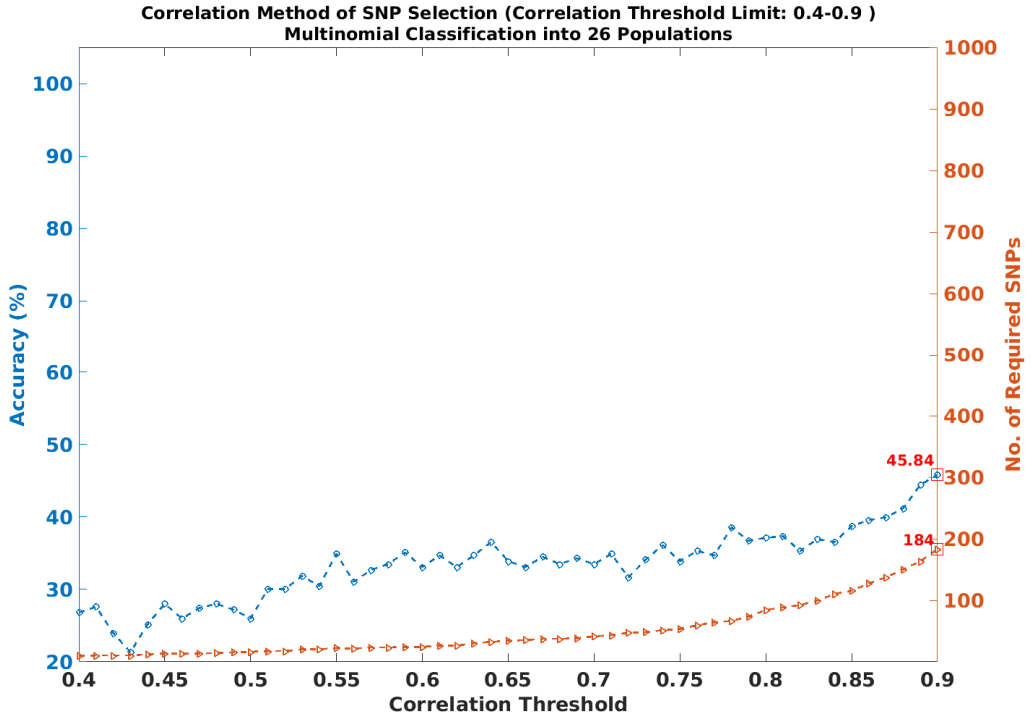
4.4.1 Multinomial Ancestry Classification Performance

In this thesis, we first introduced a correlation based algorithm for SNP selection and the experimental results obtained using this method are presented in Chapter 3. The results indicated that this method can successfully address two separate problems of ancestry classification-(i) identify a small set of SNPs for continental level ancestry classification, where continental populations are quite distant, and (ii) identify a small panel of SNPs for binary classification of any closely related sub-population pairs. However, we haven't demonstrated results on how this SNP selection method works for multi-class ancestry classification of closely associated sub-populations. We performed experiments on several cases of multi-class classification using correlation method and listed those results in the following Table 4-5 including comparison with the results obtained from random sampling method. Table 4-5 shows that in 26-class ancestry classification, correlation method requires 2477 SNPs to obtain 67.95% classification accuracy, whereas we can achieve 26-class classification accuracy of 78.50% with 1900 SNPs using the random sampling method. That is, random sampling method outperforms correlation method in 26-class ancestry classification problem. Also, from the table it is evident that in multi-class sub-continental level classification for each of the five continents, random subspace method can provide better classification accuracy with less number of SNPs in comparison to the correlation based method.

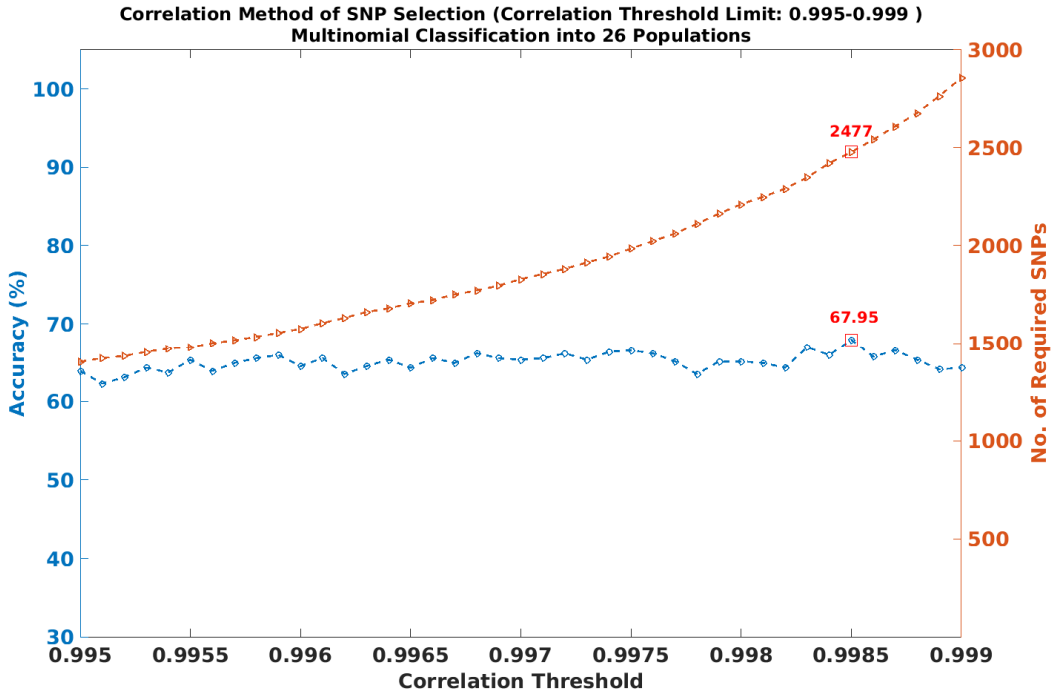
Table 4-5: Comparative Performance Analysis of Correlation Method and Random Sampling Method in Multinomial Classification

Multi-class classification	Correlation Method		Random Subspace Method	
	SNPs Required	Classification Accuracy	SNPs Required	Classification Accuracy
26 populations	2477	67.95%	1900	78.50%
5 subpopulations in Europe	1722	58.42%	1400	75.25%
4 subpopulations in America	2348	85.94%	1600	87.5%
5 subpopulations in East Asia	1855	67.33%	1800	82.18%
5 subpopulations in South Asia	1501	63.27%	1000	71.43%
7 subpopulations in Africa	1855	82.95%	1500	87.6%

We include graphical representations for some of the results mentioned in Table 4-5 in the following Figure 4-7 and Figure 4-8. In Chapter 3, we demonstrated that empirical experiments were carried out for a set of values of the correlation threshold. For each value of the correlation threshold a panel of SNPs is obtained and the learning algorithm is applied to perform classification using that particular SNP panel. For the choice of correlation threshold between 0.4 to 0.9, we obtained panels with few hundred SNPs. In both continental classification and binary classification of sub-populations, we were able to use such small panel of SNPs to perform sufficiently accurate classification. However, such small panel of SNPs is not useful while performing multinomial sub-population classification. Figure 4-7(a), depicts the performance of 26-class classification for a range of correlation threshold between 0.4 to 0.9. The highest classification performance obtained is 45.84% using 184 SNPs. Thus, we run experiments with higher values of correlation threshold in order to select more SNPs for the classification task. In Figure 4-7(b), we demonstrate the results of 26-class classification using a range of correlation thresholds between 0.995 to 0.999. It is observed that the classification accuracy now improves to 67.95% using as many as 2477 SNPs compared to 45.84% with only 184 SNPs. However, this performance result of 67.95% using 2477 SNPs is not comparable to the result obtained from random sampling method, which is 78.50% using 1900 SNPs for 26-class classification. In addition, we demonstrate the results obtained for multi-class sub-continental classification within continent Europe using the correlation method in Figure 4-8. Here it is noticed that using 1722 SNPs we can achieve classification accuracy of 58.42% and increasing SNPs beyond that cannot improve the classification performance. This classification result is also lower than the result obtained for random subspace method, which is 75.25% classification rate with 1400 SNPs.



(a)



(b)

Figure 4-7: Correlation method for 26-class classification (a) correlation threshold range: 0.4 to 0.9, (b) correlation threshold range: 0.995 to 0.999

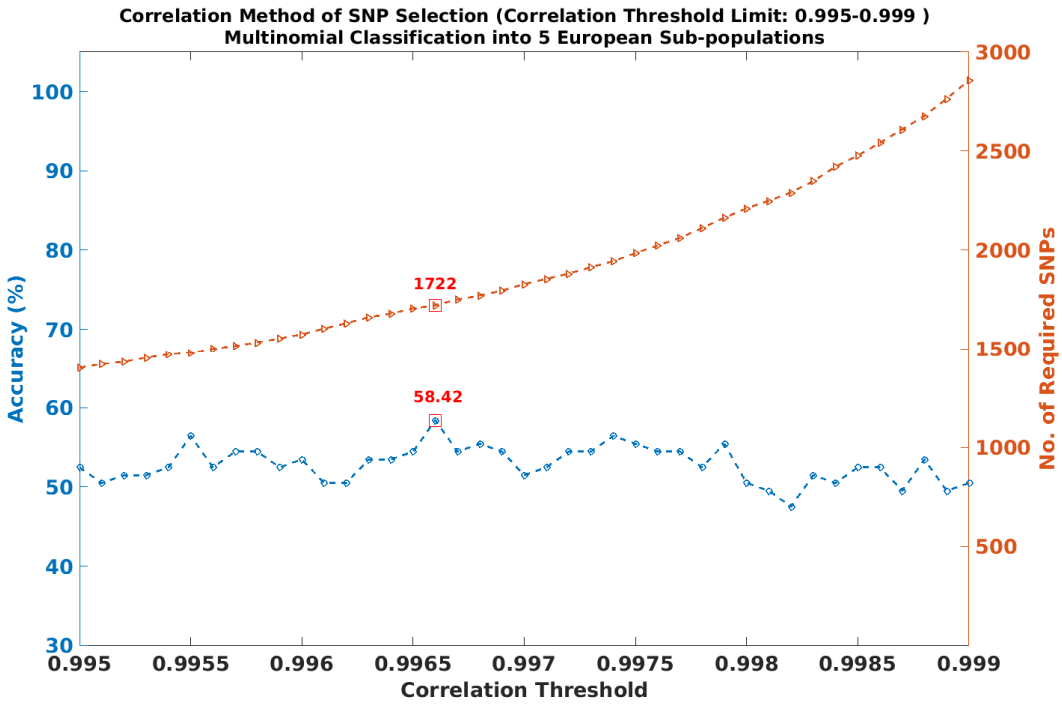


Figure 4-8: Correlation method for 5-class subcontinental classification within continent Europe (correlation threshold range: 0.995 to 0.999)

Thus, random subspace projection method for SNP selection can better address the multinomial ancestry classification problem compared to the correlation based SNP selection method.

4.4.2 Computation Time

In correlation algorithm, we evaluated the predictive power of each SNP. Each SNP has been independently used to perform ancestry classification. In Chapter 3, we notice that a performance matrix has been generated before initiating SNP selection for a certain correlation threshold. With an initial set of 6404 SNPs, the algorithm had to run 6404 times to generate the performance matrix, where each time only one SNP is being used to perform classification. The average time it takes to evaluate the performance of a SNPs is approximately 1.17 seconds. With 6404 SNPs, the time required to construct the whole performance matrix is about 2 hrs. By using a graphics processing unit (GPU), the total time of generating the performance matrix can be reduced to 1.5 hrs. Once the

performance matrix is generated, SNP selection process starts. We computed pairwise correlation between SNPs and based on a certain correlation threshold we identified a panel of non-redundant/important SNPs. Here, the value of the correlation threshold determines the size of the SNP panel and the number of SNP features in a panel determines how much time will be taken by the classifier to perform classification. For instance, SNP selection time for continental level classification using correlation threshold 0.9 is approximately 27.35 seconds, where 184 SNPs have been selected.

On the other hand, in random sampling method we select a random combination of SNPs for large number of iterations. We performed experiments with different parameters, such as size of SNP subset in each iteration, *M* and Number of iterations *N*. It is found that with *M*=50, the algorithm takes an average of 8.13 seconds for random selection of 50 SNPs and performing classification. If the algorithm is run for 50000 iterations (*N*=50000), it approximately takes 112.95 hours (≈ 4.7 days) to complete the selection of random subsets. Using GPU, the overall time can be reduced to 69.21 hours (≈ 2.8 days). Once the random subsets are selected from many iterations, the top *Q* subsets are chosen and each SNP in the top *Q* subsets is assigned an individual ranking based on their number of occurrence in the top *Q* subsets. Based on the choice of Q, the required time to rank the SNPs differs. For example, for *Q* =10000 the overall time required to rank the random subsets and individual SNP from top *Q* subsets is approximately 1.7 hrs. The comparative analysis on computation time at different stages of the two algorithms is shown in Table 4-6.

Overall, it is observed that the correlation method requires much less time during the process of SNP selection in comparison to the random subspace method. However, once we obtain a specific panel of SNPs for distinguishing between a certain group of populations using either of the two methods, the time required to perform classification relies only on the number of SNP features selected.

Table 4-6: Computation Time during Algorithm Construction

Correlation Method	Random Sampling Method
Individual SNP performance ≈ 1.17 seconds Overall 6404 SNPs ≈ 2 hrs (1.5 hrs with GPU)	M=50 SNPs Performance ≈ 8.13 seconds Overall N=50000 iterations ≈ 4.7 days (2.8 days with GPU)
SNP selection time for a certain correlation threshold (th=0.9) ≈ 27.35 seconds	SNP ranking time for Q=10000 (ranking random subsets + ranking individual SNP) ≈ 1.7 hrs

4.5 Conclusions

In this chapter, we have designed a SNPs selection algorithm exploiting random subspace projection approach. This approach has been observed to be very effective in selecting small subsets of ancestry informative SNPs for distinguishing multiple closely associated sub-populations in the same continent. We noticed that sub-populations within continent America, East Asia and Africa are relatively easy to distinguish, whereas more difficulties arise while distinguishing between the sub-populations within South Asia and sub-populations within Europe. We could further increase the performance of our overall two-stage ancestry estimation model if we could perform better in within continent multi-class classification for these two continents. Moreover, it is observed that in the multinomial ancestry classification of sub-populations, random sampling method provides significantly better performance in comparison to the correlation based method despite taking longer time during the SNP selection process.

Chapter 5: Conclusion and Future Work

In this work, we have addressed continental and sub-continental ancestry estimation problems in a resource constrained environment. We analyzed only the DNA of Chromosome 1, which is the largest human chromosome, to identify the ancestry informative marker SNPs. In order to develop an ancestry estimation model, we performed SNP selection in multiple stages. In the initial stages of selection, we first applied a parameter based selection. Next, further pruning has been conducted in the outlier based selection stage using the DBSCAN clustering algorithm. Later, we developed two different approaches for final stage of SNP selection. In one approach, we applied a correlation based filtering method, where pairwise correlation of SNPs is computed to remove the redundant SNPs from the analysis. In this approach, we have evaluated the discriminant power of each SNP individually and used the individual performance metric to calculate pairwise correlation between SNPs. With the choice of a correlation threshold, some SNPs appeared to be redundant and removed from the analysis. We applied this correlation based filtering technique to identify the important SNPs for continental level classification as well as binary classification between closely related sub-populations. Here, once the relevant SNPs are identified, ancestry classification is performed on the test set using the softmax neural network classifier. The continental classification accuracy using the correlation based approach is as high as 96.75% using 206 SNPs and it can reach up to 99.19% using 614SNPs. The binary/pairwise classification performances between the sub-populations are sufficiently high in most cases using a few marker SNPs. In a number of cases of binary sub-population classification, we achieved 100% classification accuracy, such as African sub-populations Gambian vs. Luhya, South Asian sub-populations Punjabi vs. Gujarati, American sub-populations Puerto Rico vs. Peru. But, also there are several challenging cases with binary classification rates in range of 60%-70%, for instance, Puerto Rico vs. Columbia in America, Sri

Lankan vs. Indian in South Asia. Apart from the correlation based approach, the other SNPs selection approach is based on random subspace projection. This is an iterative feature selection technique, which considers potential interactions among the SNPs in the random subspace. Here, the learning algorithm, softmax neural network, itself evaluates the usefulness of SNPs features and removed the noisy ones. SNPs have been identified for both continental-level classification and sub-continental level multi-class classification using this approach. For this approach, we can achieve continental accuracy of 97.57% using 210 SNPs and this performance can be improved further up to 99.19% using 380 SNPs. In case of multi-class classification of closely related subpopulations, we also achieved sufficiently good classification rate using less than 2000 SNPs. For instance, multi-class classification accuracy between seven closely related African sub-populations is as high as 87.6% using 1500 SNPs. Also, similar performance achieved while distinguishing four American sub-populations. But, distinguishing the sub-populations in South Asia is relatively difficult with achieved multinomial classification rate of 71.43%. Finally, using the continent informative SNPs and sub-continental informative SNPs obtained through executing the random subspace projection algorithm, we have developed a two-step ancestry prediction model. This two-step model predicts an individual's exact ancestry by first predicting the continental origin and later predicting the sub-population identity by comparing between the sub-populations within the detected continent. The random subspace projection technique took much longer time to identify the best informative SNPs compared to the correlation based technique, since the learning algorithm is called repeatedly in that approach. However, this approach demonstrated superior performance in the difficult multinomial sub-population classification.

Along this line of research, possible direction for future steps can be listed as follows:

- In this study, we focused on Chromosome 1 to infer ancestry. In future, we need to analyze other chromosomes using our proposed methods and investigate whether any other chromosome contain better marker SNPs for ancestry estimation.
- We have not conducted our analysis on admixed populations. Separating admixed populations are challenging mostly when the ancestral populations are closely related. As a future work, we can apply our proposed technique on an admixture dataset.
- In our SNP selection methods, we have not considered the impact of linkage disequilibrium. As we know that, genes which are in linkage disequilibrium might contain SNPs of similar allele information. In future, we plan to refine our SNP selection methods by ignoring the SNPs from the closely located genes which are in linkage disequilibrium.
- We performed an outlier based SNP selection using DBSCAN clustering in the initial stage of SNPs pruning. However, as a future work we may plan to investigate whether the cluster centroids perform as better markers instead of the outliers.
- We used random subspace projection technique particularly for multi-class ancestry classification problems, that is, multi-class continental classification and multi-class sub-continental classification. We can apply this technique to solve several difficult cases of pairwise sub-population classifications.

References

1. Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D: “Using ancestry informative markers to define populations and detect population stratification,” J Psychopharmacol 2006, 20:199-126.
2. Araújo, Gilderlanio S., et al. "Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks)." Bioinformatics 32.8 (2016): 1247-1249.
3. Bhaskar, Anand, Adel Javanmard, Thomas A. Courtade, and David Tse. "Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies." Bioinformatics 33, no. 6 (2016): 879-885.
4. Fondevila, M., et al. "Revision of the SNPforID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies." Forensic Science International: Genetics 7.1 (2013): 63-74.
5. Gettings, Katherine Butler, et al. "A 50-SNP assay for biogeographic ancestry and phenotype prediction in the US population." Forensic Science International: Genetics 8.1 (2014): 101-108.
6. Tian, Chao, et al. "A genomewide single-nucleotide–polymorphism panel for Mexican American admixture mapping." The American Journal of Human Genetics 80.6 (2007): 1014-1023.
7. Sanderson, Jean, et al. "Reconstructing past admixture processes from local genomic ancestry using wavelet transformation." Genetics 200.2 (2015): 469-481.
8. Arthur, Rudy, et al. "AKT: ancestry and kinship toolkit." Bioinformatics 33.1 (2017): 142-144.
9. Krimsky, S and Simoncelli, T, Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties, Columbia University Press, (2012).
10. Ragna Aarli, "Genetic Justice and Transformations of Criminal Procedure" Journal of Scandinavian Studies in Criminology and Crime Prevention 13:1, (2012), 3-21.
11. Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., ... & Carracedo, Á. (2009). Ancestry analysis in the 11-M Madrid bomb attack investigation. PLoS One, 4(8), e6583.
12. Wen, Wanqing, Xiao-ou Shu, Xingyi Guo, Qiuyin Cai, Jirong Long, Manjeet K. Bolla, Kyriaki Michailidou et al. "Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry." Breast Cancer Research 18, no. 1 (2016): 124.
13. Bandera, Elisa V., Urmila Chandran, Gary Zirpoli, Zhihong Gong, Susan E. McCann, Chi-Chen Hong, Gregory Ciupak, Karen Pawlish, and Christine B. Ambrosone. "Body fatness and breast cancer risk in women of African ancestry." BMC Cancer 13, no. 1 (2013): 475.
14. Phillips, C., A. Salas, J. J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza et al. "Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs." Forensic Science International: Genetics 1, no. 3 (2007): 273-280.
15. Kidd, Judith R., et al. "Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples." Investigative Genetics 2.1 (2011): 1.
16. Nassir, Rami, et al. "An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels." BMC Genetics 10.1 (2009): 39.
17. Paschou, Peristera, Elad Ziv, Esteban G. Burchard, Shweta Choudhry, William Rodriguez-Cintron, Michael W. Mahoney, and Petros Drineas. "PCA-correlated SNPs for structure identification in worldwide human populations." PLoS Genetics 3, no. 9 (2007): e160.

18. Liu, Yushi, Toru Nyunoya, Shuguang Leng, Steven A. Belinsky, Yohannes Tesfaigzi, and Shannon Bruse. "Softwares and methods for estimating genetic ancestry in human populations." Human genomics 7, no. 1 (2013): 1.

19. Pardo-Seco, Jacobo, Federico Martinón-Torres, and Antonio Salas. "Evaluating the accuracy of AIM panels at quantifying genome ancestry." BMC Genomics 15, no. 1 (2014): 543.

20. Wright S: Evolution and the Genetics of Populations, vol 2: The Theory of Gene Frequencies Chicago and London: University of Chicago Press; 1969.

21. Price, Alkes L., et al. "Discerning the ancestry of European Americans in genetic association studies." PLoS genetics 4.1 (2008): e236.

22. Mao, Xianyun, et al. "A genomewide admixture mapping panel for Hispanic/Latino populations." The American Journal of Human Genetics 80.6 (2007): 1171-1178.

23. Seldin, Michael F., et al. "European population substructure: clustering of northern and southern populations." PLoS Genetics 2.9 (2006): e143.

24. Campbell, Catarina D., et al. "Demonstrating stratification in a European American population." Nature Genetics 37.8 (2005): 868.

25. Seldin, Michael F., and Alkes L. Price. "Application of ancestry informative markers to association studies in European Americans." PLoS Genetics 4.1 (2008): e5.

26. Tian, Chao, et al. "Analysis of East Asia genetic substructure using genome-wide SNP arrays." PLoS One 3.12 (2008): e3862.

27. Bryc, Katarzyna, et al. "Genome-wide patterns of population structure and admixture in West Africans and African Americans." Proceedings of the National Academy of Sciences 107.2 (2010): 786-791.

28. Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature Genetics 38.8 (2006): 904.

29. Novembre, John, and Matthew Stephens. "Interpreting principal component analyses of spatial population genetic variation." Nature Genetics 40.5 (2008): 646-649.

30. Patterson, Nick, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." PLoS Genetics 2.12 (2006): e190.

31. Byun, Jinyoung, Younghun Han, Ivan P. Gorlov, Jonathan A. Busam, Michael F. Seldin, and Christopher I. Amos. "Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure." BMC Genomics 18, no. 1 (2017): 789.

32. Li Y, Byun J, Cai G, et al, “FastProp: A rapid pricipal componet dericed method to infer intercontenental ancestry using genetic data”, BMC Bioinformatics, 17:122 (2016).

33. Pritchard, Jonathan K., et al. "Association mapping in structured populations." The American Journal of Human Genetics 67.1 (2000): 170-181.

34. Kidd, Kenneth K., et al. "Progress toward an efficient panel of SNPs for ancestry inference." Forensic Science International: Genetics 10 (2014): 23-32.

35. Baran, Yael, et al. "Fast and accurate inference of local ancestry in Latino populations." Bioinformatics 28.10 (2012): 1359-1367.

36. Chimusa, Emile R., et al. "ancGWAS: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations." Bioinformatics 32.4 (2016): 549-556.

37. 1000 Genomes Project Consortium. "A global reference for human genetic variation." Nature 526.7571 (2015): 68.

38. Nachman, Michael W. "Single nucleotide polymorphisms and recombination rate in humans." TRENDS in Genetics 17, no. 9 (2001): 481-485.

39. Gymrek, Melissa, Thomas Willems, David E. Reich, and Yaniv Erlich. "A framework to interpret short tandem repeat variation in humans." bioRxiv (2016): 092734.

40. Chen, Yu-Sheng, Antonio Torroni, Laurent Excoffier, A. Silvana Santachiara-Benerecetti, and Douglas C. Wallace. "Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups." American journal of human genetics 57, no. 1 (1995): 133.

41. Shriver, Mark D., Rui Mei, Esteban J. Parra, Vibhor Sonpar, Indrani Halder, Sarah A. Tishkoff, Theodore G. Schurr et al. "Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation." Human Genomics 2, no. 2 (2005): 81.

42. Kosoy, Roman, Rami Nassir, Chao Tian, Phoebe A. White, Lesley M. Butler, Gabriel Silva, Rick Kittles et al. "Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America." Human Mutation 30, no. 1 (2009): 69-78.

43. Lins, Tulio C., Rodrigo G. Vieira, Breno S. Abreu, Dario Grattapaglia, and Rinaldo W. Pereira. "Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs." American Journal of Human Biology 22, no. 2 (2010): 187-192.

44. Hashiyada, M., Y. Itakura, T. Nagashima, M. Nata, and M. Funayama. "Polymorphism of 17 STRs by multiplex analysis in Japanese population." Forensic Science International 133, no. 3 (2003): 250-253.

45. Graydon, Matthew, François Cholette, and Lay-Keow Ng. "Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits." Forensic Science International: Genetics 3.4 (2009): 251-254.

46. Londin, Eric R., Margaret A. Keller, Cathleen Maista, Gretchen Smith, Laura A. Mamounas, Ran Zhang, Steven J. Madore, Katrina Gwinn, and Roderick A. Corriveau. "CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins." PLoS One 5, no. 10 (2010): e13443.

47. Silva, Nuno M., Luísa Pereira, Estella S. Poloni, and Mathias Currat. "Human neutral genetic variation and forensic STR data." PLoS One 7, no. 11 (2012): e49666.

48. Corach, Daniel, Oscar Lao, Cecilia Bobillo, Kristiaan van Der Gaag, Sofia Zuniga, Mark Vermeulen, Kate Van Duijn et al. "Inferring continental ancestry of Argentineans from autosomal, Y-chromosomal and mitochondrial DNA." Annals of Human Genetics 74, no. 1 (2010): 65-76.

49. Rishishwar, Lavanya, Andrew B. Conley, Brani Vidakovic, and I. King Jordan. "A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians." Gene 574, no. 2 (2015): 345-351.

50. Nievergelt, Caroline M., Adam X. Maihofer, Tatyana Shekhtman, Ondrej Libiger, Xudong Wang, Kenneth K. Kidd, and Judith R. Kidd. "Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel." Investigative genetics 4, no. 1 (2013): 13.

51. Weir, Bruce S., and C. Clark Cockerham. "Estimating F-statistics for the analysis of population structure." evolution 38, no. 6 (1984): 1358-1370.

52. Li, Yafang, Jinyoung Byun, Guoshuai Cai, Xiangjun Xiao, Younghun Han, Olivier Cornelis, James E. Dinulos et al. "FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data." BMC Bioinformatics 17, no. 1 (2016): 122.

53. Hajiloo, Mohsen, et al. "ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction." BMC Bioinformatics 14.1 (2013): 61.

54. Sankararaman, Sriram, Srinath Sridhar, Gad Kimmel, and Eran Halperin. "Estimating local ancestry in admixed populations." The American Journal of Human Genetics 82, no. 2 (2008): 290-303.

55. Yang, James J., Jia Li, Anne Buu, and L. Keoki Williams. "Efficient inference of local ancestry." Bioinformatics 29, no. 21 (2013): 2750-2756.

56. Paşaniuc, Bogdan, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. "Inference of locus-specific ancestry in closely related populations." Bioinformatics 25, no. 12 (2009): i213-i221.

57. Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3, no. Mar (2003): 1157-1182.

58. Neumann, Julia, Christoph Schnörr, and Gabriele Steidl. "Combined SVM-based feature selection and classification." Machine Learning 61, no. 1 (2005): 129-150.

59. Song, Guo-Jie, Shi-Wei Tang, Dong-Qing Yang, and Teng-Jiao Wang. "A spatial feature selection method based on maximum entropy theory." Journal of Software 14, no. 9 (2003): 1544-1550.

60. Last, Mark, Abraham Kandel, and Oded Maimon. "Information-theoretic algorithm for feature selection." Pattern Recognition Letters 22, no. 6 (2001): 799-811.

61. Hall, Mark A., and Lloyd A. Smith. "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper." In FLAIRS conference, vol. 1999, pp. 235-239. 1999.

62. Whitley, David C., Martyn G. Ford, and David J. Livingstone. "Unsupervised forward selection: a method for eliminating redundant variables." Journal of Chemical Information and Computer Sciences 40, no. 5 (2000): 1160-1168.

63. Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." KDD. 96.34 (1996).

64. Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.

65. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

66. Lao, Oscar, et al. "Evaluating self-declared ancestry of US Americans with autosomal, Y-chromosomal and mitochondrial DNA." Human Mutation 31.12 (2010).

67. Teisseyre, P., Kłopotek, R. A., & Mielniczuk, J. (2016). Random Subspace Method for high-dimensional regression with the R. Computational Statistics, 31(3), 943-972.

68. Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.

69. Li, X., & Zhao, H. (2009). Weighted random subspace method for high dimensional data classification. Statistics and Its Interface, 2(2), 153–159.

70. Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random KNN feature selection-a fast and stable alternative to Random Forests. BMC Bioinformatics, 12(1), 450.

71. Lai, C., Reinders, M. J., & Wessels, L. (2006). Random subspace method for multivariate feature selection. Pattern Recognition Letters, 27(10), 1067-1076.

72. Cai, R., Hao, Z., & Wen, W. (2007, August). A novel gene ranking algorithm based on random subspace method. In Neural Networks, 2007. IJCNN 2007. International Joint Conference on (pp. 219-223). IEEE.

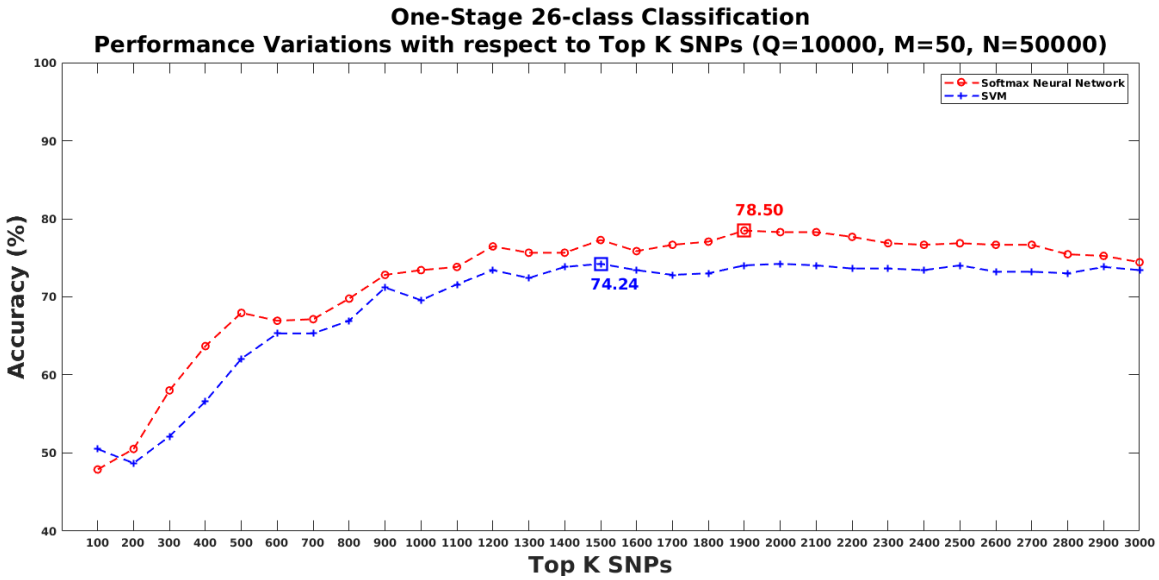
Appendix

A: Neural Network vs. SVM for Ancestry Classification

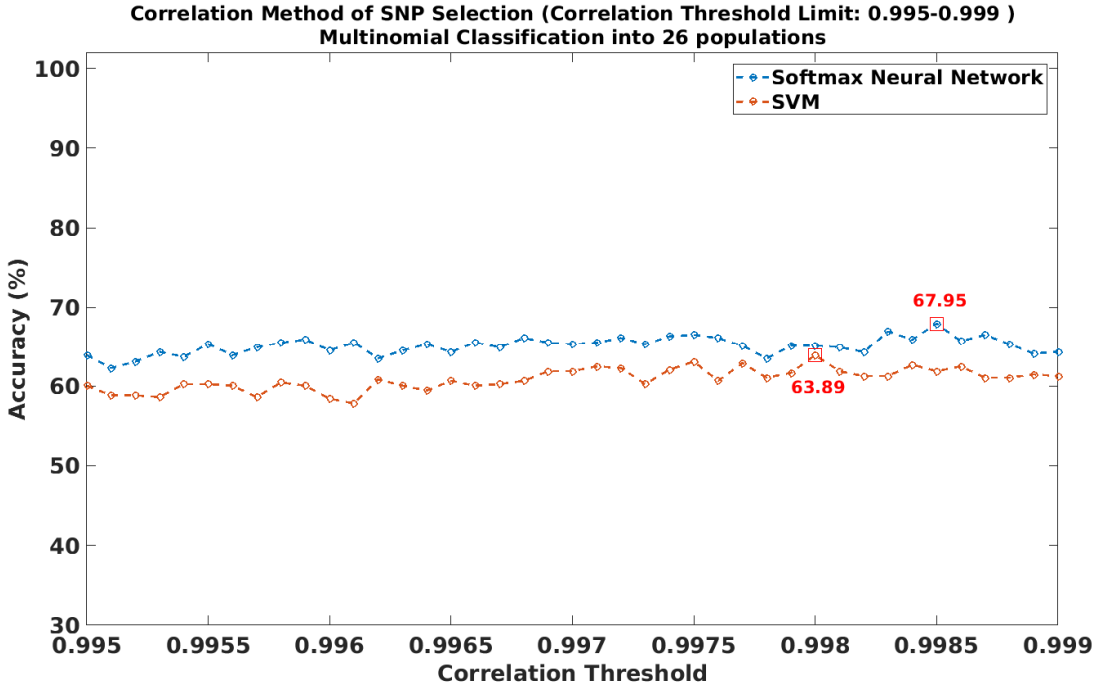
Artificial Neural Networks is a biologically inspired network of artificial neurons configured to perform specific tasks. Nowadays, neural network architectures are performing significantly better than other learning algorithms in solving complex non-linear hypothesis due to the surge of training data and faster computers. A neural network learns its own features, that is, the features at the hidden layer themselves are learned as the function of the inputs. The original features from the input layer are mapped into more complex features in the hidden layer, thus eventually yields better hypothesis, better prediction.

Support vector machine (SVM) is another very popular supervised learning algorithm, which can solve the local minima problem and overfitting issues that might be encountered by neural network architectures. However, in many applications where the size of training data is very large, neural network can outperform SVM or other logistic regression classifiers.

In the problem of ancestry classification studied in this thesis, we observe that neural network architecture with softmax activation at the output layer performs better in the classification stage compared to the SVM classifier. In the following Figure A-1 (a & b), the classification performance of 26-class ancestry classification is demonstrated for both approaches-correlation method and random sampling method. In the classification stage of both methods, we observe that neural network classifier outperforms the SVM classification scheme.



(a)



(b)

Figure A-1: (a) 26-class classification performances over top *K* SNPs using random sampling method, softmax neural network performance vs. SVM performance (b) 26-class classification performances over a range of correlation thresholds using correlation method, softmax neural network performance vs. SVM performance